# Classifying Canadian citizens' FWB status and predicting the impact of global shocks using machine learning

Supervisor: Prof. Ahmed El-Roby
School of Computer Science, Carleton University
ahmedelroby@cunet.carleton.ca

Aziz Al-Najjar
Department of Systems and Computer Engineering, Carleton University
azizalnajjar@cmail.carleton.ca

Huiqian Chen
Sprott School of Business, Carleton University
huiqianchen@cmail.carleton.ca

Zakaria Zoundi
Department of Economics, Carleton University
zakariazoundi@cmail.carleton.ca

## ABSTRACT

This project utilizes FWB data from both pre- and post-COVID-19 surveys conducted by the Financial Consumer Agency of Canada (FCAC) to develop a machine learning model that can predict Canadians' FWB status and the potential impact of future global shocks. The project aims to investigate whether the key factors contributing to FWB change when ignoring COVID-19, and whether the same set of features can accurately predict FWB in both pre- and post-COVID-19 datasets. The results show that FWB is disproportionately distributed across the nation. The key drivers of FWB include: household financial situation, capacity to meet monthly expenses, saving, credit score. Besides, the analysis showed that the prediction of FWB has a higher level of accuracy when global shocks are not accounted for. This showed the significant disruptive effects the pandemic induced across sectors and individuals. The project advocates for an increasing effort from the government to help Canadians across provinces to better cope with their finances. Addressing this issue involves creating better jobs, a review of minimum wages, controlling inflation, supporting debt relief including mortgage, and further assistance to households in the lowest income deciles, among others. Particular attention should be given to groups that are usually left-behind such as women and single mothers, persons living with disability and indigenous communities.

## 1 INTRODUCTION

Financial well-being (FWB) is crucial in an individual's life, connecting their financial satisfaction with their current and future situations. This project uses pre- and post-COVID-19 data from the Financial Consumer Agency of Canada (FCAC) to develop a machine learning model for predicting Canadians' FWB and potential responses to future global crises. The project addresses critical questions about FWB, such as evaluation methods, prediction accuracy, and the stability of influencing factors across different periods. The insights gained can help identify at-risk populations and assist decision-makers in creating targeted support programs to mitigate the adverse effects of global shocks on FWB.

In the pioneering research on FWB classification, the main focus has been on defining FWB, the necessity of research on the topic, and its measurement methods. However, no consensus exists on these aspects. Campbell[5] laid the foundation for numerous FWB studies across various fields, including financial guidance and planning [2], budget management, consumer decision-making, service marketing, and responsible financing in the fintech era [11]. Classifying FWB is a complex task, as previous researchers have used either positive paradigms (such as financial satisfaction) or negative paradigms (such as financial stress) to identify it. Presently, a comprehensive classification of FWB requires a three-dimensional approach [9]: subjective, objective, and psychological, which includes a set of measures for evaluation. Building on the spectrum of areas covered by these dimensions, researchers have suggested the classification of individuals' FWB in a more heterogeneous way, using some benchmarks or score, rather than relying on a single metric [3]. The United Sates Consumer Financial Protection Bureau, for example, applied this concept and note that on average, FWB of Americans adults increased slightly between 2017 and 2020. However, over the period, between 10% and 13% of Americans had a low level of FWB and less than 20% a very high FWB [1].

The ongoing debate emphasizes the importance of further exploration of the measures and categorization of individual scores, and provides valuable insights into the study of the FWB, particularly for countries like Canada in the wake of the COVID-19 pandemic.

The structure of this paper is as follows: Section 2 provides a description of the datasets used in this project, while Section 3 presents the methodology adopted to address the problem, including data preprocessing, the description of the selected model, and an explanation of how the model was implemented. In Section 4, the results and findings on factors influencing FWB and stability across different models are discussed. Section 5 analyzes possible reasons behind the results reported by the three models and discusses the key factors and key points. Finally, in Section 6, the study's results and future directions are summarized.

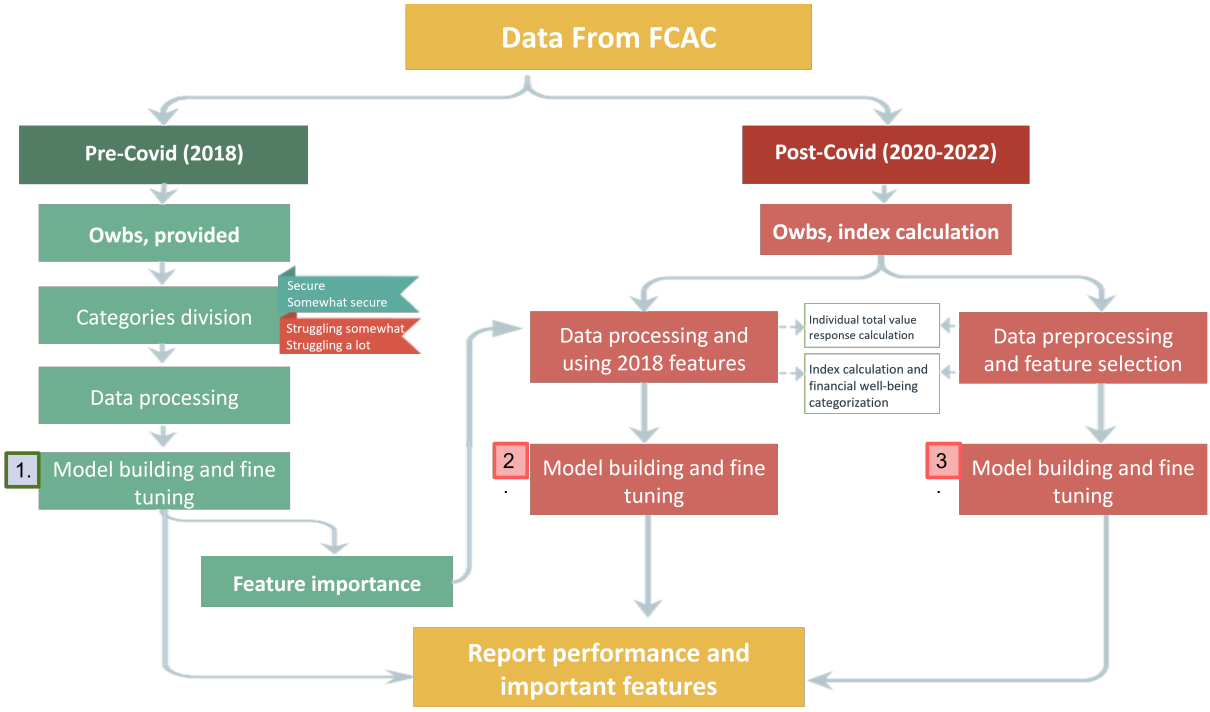


**Figure 1: Overall Project methodology**

## 2 DATASET DESCRIPTION

Data collection presented one of the core aspects of this research, since they drove the entire analysis, results and policy recommendation. After a series of research and consultations, the project finally used two main datasets. The first dataset is the 2018 FWB and Capability [6] with 1935 observations. The second dataset is the 2020-22 COVID-19 FWB survey, which contains 9,394 observations [7].

As the names of the surveys suggest, the first dataset is related to the period before COVID-19, while the second was launched in the context of the pandemic. Both datasets were provided by the Financial Consumer Agency of Canada (FCAC) upon request. The surveys are parts of the Canadian Financial Capability Survey (CFCS) which aim at shedding light on Canadians' knowledge, capacities, and confidence regarding financial decision-making. The 2020-22 COVID-19 FWB Survey, which constitutes the backbone of the analysis (due to the questions specific to COVID-19) covers the period August 2020 to September 2022. A total of 9,394 interviews targeted toward Canadians aged 18 years or older were set up. The components of the survey, which are also partly included in the 2018 survey, incorporate data on: socio-demographics, labor market participation and income; ongoing-expenses and day-to-day financial management of the household; assets, debts and credit management; paying down debt and setting aside money for an emergency fund; psychological characteristics and attitudes towards money; change in respondents' situation due to COVID-19; FWB.

## 3 METHODOLOGY

Figure 1 displays the project plan, which involves leveraging two datasets provided by the FCAC and developing three distinct models. The first model will be a pre-COVID 19 model that uses the FCAC 2018 dataset. This dataset includes a third-party-calculated Overall Well-Being Score (OWBS) for each respondent in addition to the survey questions and answers. To categorize the subjects, we will use the four FWB categories recommended by the FCAC, which are determined based on the OWBS. Specifically, we will divide the subjects into the following categories:

- OWBS from 0 to 30 : Struggling a Lot
- OWBS from 31 to 50: Struggling Somewhat
- OWBS from 51 to 80: Somewhat Secure
- OWBS from 81 to 100: Secure

To prepare the data for model development, the data will undergo multiple preprocessing steps, which will be elaborated on in Section 3.1. Next, we will construct a machine-learning model and fine-tune it to achieve the highest possible accuracy. For more information on the chosen machine-learning model, see Section 3.2. The machine learning model will determine the importance of each feature, or survey question, that contributed to the classification, which will be used to develop the second model.

Moving on to the second and third models, they will use the post-COVID 19 survey dataset provided by the FCAC. Unlike the 2018 survey, the OWBS for each respondent was not provided in this dataset. However, with the help of FCAC, we calculated the OWBS ourselves, as described in Section 3.3. We then categorized the respondents into one of the four FWB categories outlined earlier.

For the second model, we will only utilize the 30 most important questions identified by the first model. The purpose of this model is to provide insights into the FWB of the population after the pandemic, without considering the COVID-related questions. Moreover, we will investigate whether we can reach similar performance using only the top pre-COVID questions.

Finally, the third model will employ all the questions in the post-COVID 19 dataset without linking them to any pre-COVID features. The data will still undergo the same preprocessing and feature selection process as the other models.

## 3.1 Data preprocessing

The datasets we are working with capture the financial behavior of Canadians, with the 2020 dataset containing additional questions related to COVID-19 that can help in studying the impact of the pandemic on FWB. However, before we can start training our models, we need to preprocess and select the most important features to ensure that the data is ready for analysis, as illustrated in Figure 3. The following steps were taken:

(1) Drop uninformative columns:
We removed columns such as response ID and time of the phone call, which do not provide any useful information for the analysis.
(2) Replace empty cells with NaN:
Missing values can be informative and provide insight into patterns of missing data, so we replaced empty cells with NaN values.
(3) Encoding string columns using label encoding:
Since machine learning models typically require numerical input, we encoded string columns using label encoding. This assigns a numerical value to each category, enabling the model to process the data.
(4) Rewrite numeric answers and ensure they are in floats:
Some answers were recorded as strings rather than numbers, so we rewrote and converted them to floats to ensure consistency and that the data could be processed by the model.
(5) Handling data imbalance by resampling:
The dataset classes were imbalanced, with some classes having significantly fewer examples than others, as can be seen in 2. To address this, we performed resampling to increase the number of examples in the minority classes. This involved randomly duplicating some examples from the minority classes until they had the same number of examples as the majority class. However, it is important to note that this step was only performed on the training dataset and not on the testing set.
(6) Dropping questions used in calculating the well-being index:
Since the well-being index is a composite score calculated from several questions, including these questions in the analysis would lead to data leakage, where the model learns to predict the target variable based on the questions used to calculate it. Therefore, we dropped these questions from the dataset to avoid this issue.
(7) Selecting the most correlated features with the classes:
To improve the performance of the model, we selected the features that were most strongly correlated with the target variable. This involved calculating the correlation coefficient between each feature and the target variable and selecting the top features based on this measure.

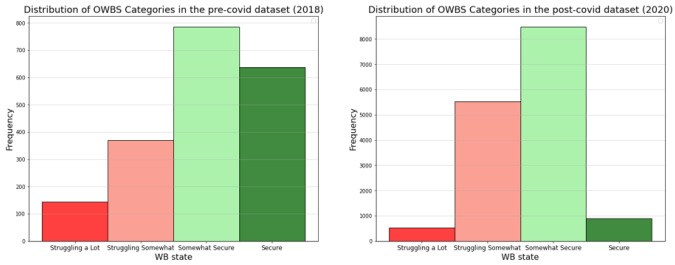


**Figure 2: Datasets categories distribution: Left histogram shows the 2018 survey distribution. Right histogram shows the post-COVID survey distribution**

These preprocessing and feature selection steps were necessary to ensure that the data was ready for analysis and to improve the performance of the model. By following these steps, we can be confident that the data we are using is clean, informative, and well-optimized for machine learning analysis

## 3.2 Model Description

In the context of our project, XGBoost was selected because of its ability to handle missing data effectively [8]. Unlike random or measurement errors, missing values in our datasets were often informative or more informative than other values, as exemplified by the 2018 data. The responses to the initial question "Do you live with any children?" would determine whether the subsequent follow-up questions on the number and ages of children would be relevant or not. Thus, traditional methods of imputing missing values like mean, median, or mode were unsuitable for our dataset.

XGBoost is a tree-based gradient boosting algorithm that is well-suited to handle missing data in tabular data [13]. It uses decision trees as weak learners and builds each tree iteratively, one at a time. At each iteration, the algorithm focuses on the incorrectly classified samples from the previous iteration and builds a new tree to correct these errors by adding them to the existing ensemble.

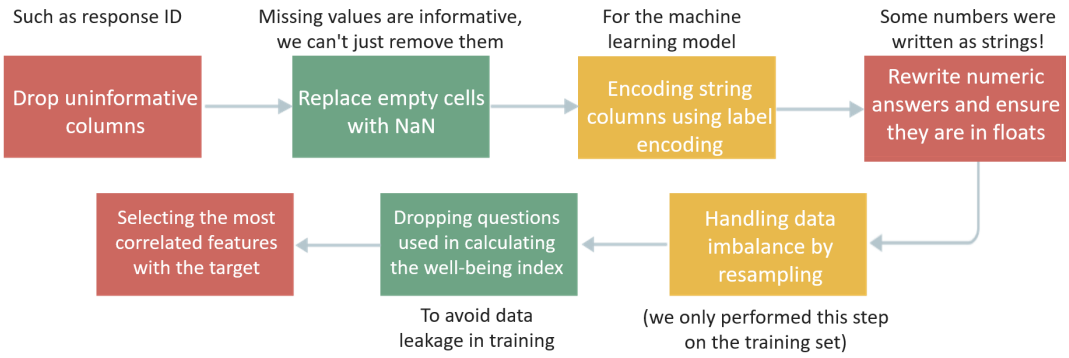


**Figure 3: Data Pre-Processing**

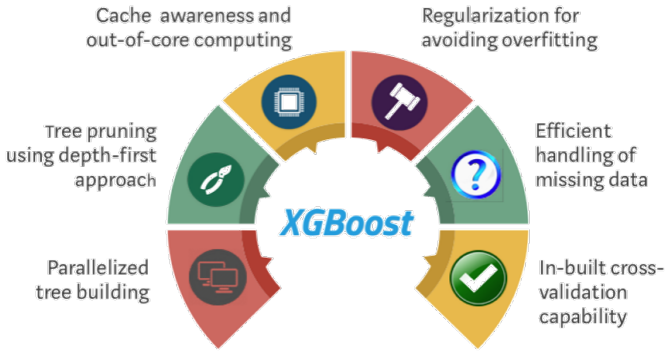


**Figure 4: XGBoost Attributes**

To prevent overfitting, XGBoost uses regularization techniques such as L1 and L2 regularization, and gradient-based one-side sampling to handle missing data. This technique allows XGBoost to learn from the presence or absence of features, making it suitable for datasets with numerous missing values. Additionally, XGBoost uses shrinkage to reduce the impact of individual trees in the ensemble, which helps to prevent overfitting and improve the model's generalization ability. More into the XGboost attributes can be found in Figure. 4

The high level of customization offered by XGBoost allows for the optimization of hyperparameters to enhance model performance. The use of a grid search algorithm to find the optimal combination of hyperparameters for each dataset's model and the evaluation of the model's performance using several metrics, including accuracy, precision, recall, and F1-score, ensuring the best possible performance of the model on our dataset.

### 3.3 Well-being index calculations

The FWB index was available only for the 2018 survey dataset. Following the suggestion from the data providers, the COVID-19 index was computed following the same concept of scoring as the 2018 data, and in line with the US Consumer Financial Protection Bureau [4]. The basic steps are depicted in the flow chart below and can be summarized as follows:

(1) Since the estimation method has two approaches, a shorter and longer version, depending on the availability of data, we began by selecting the shorter version to match the data.
(2) Keep track of responses on age and the survey mode (i.e "online / self-administered" or "phone/ administered by someone else"). Both age and survey mode are used to weight individuals' scores.
(3) Calculate each individual's total value response based on the points attributed to their response to each question related to FWB measures.
(4) Compute the FWB score (0-100), taking into account the age and survey mode. The scoring sheet (available in the methodology document), provides a straightforward way to obtain the weighted score.
(5) Based on the total score, each respondent is categorized as struggling a lot (when the well-being score ranges from 0 to 30); Struggling somewhat (30.01 to 50); Somewhat secure (50.01 to 70) and Secure (70.01 to 100).

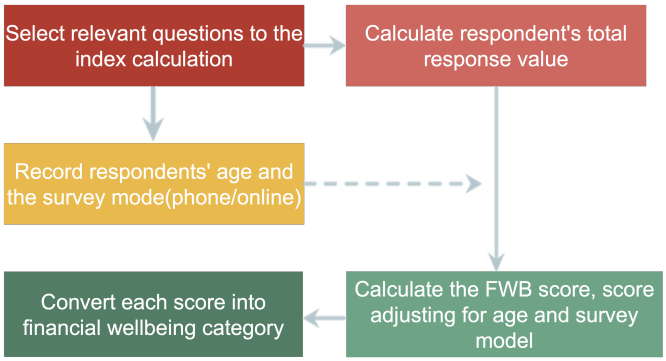


**Figure 5: FWB index calculation**

After computing the scores, the final scores and classifications were added as new columns to the existing individual features. The entire new dataset was then ready for analysis.

## 4 RESULTS

In the methodology section, we described the three models used for training, namely the pre-COVID 19 model, the post-COVID 19 model with no feature matching, and the post-COVID 19 model with feature matching. Following the pre-processing steps outlined in the methodology, we utilized the XGBoost algorithm to predict the target variable "owbs_category". To prepare the categorical target variable "owbs_category" for modeling, we used the LabelEncoder method from scikit-learn [12] to convert it into numerical labels ranging from 0 to 3, where 0 represented the 'Secure' category, and 3 represented the 'Struggling a Lot' category. In addition, we employed the SelectKBest [12] method from scikit-learn to select the most relevant features for each model, with k=35 for the pre-COVID data and k=110 for the post-COVID models. The f_regression function was used to evaluate the correlation between the features and the target variable. Subsequently, we conducted a grid search [10] to determine the optimal hyperparameters for each model. We tuned learning_rate, max_depth, n_estimators, min_child_weight, subsample, colsample_bytree, gamma, and scale_pos_weight hyperparameters as they have a significant impact on the performance of XGBoost. The best hyperparameters were then utilized to train the final models. We also employed a 5-fold cross-validation technique, where the dataset was randomly split into training and testing sets, with 70% of the data allocated for training and 30% for testing.

Upon evaluating the models on the testing set, we reported their precision, recall, f1-score, and accuracy in Table 1. As expected, the pre-COVID survey-based model exhibited superior performance in all metrics in comparison to the two post-COVID models. This could be attributed to the differences in the structures of the questions used in the post-COVID survey compared to the pre-COVID survey. Moreover, it is possible that the pre-COVID dataset had more relevant or informative features compared to the post-COVID datasets, which made it easier for the model to predict FWB accurately. Additionally, the well-being score calculated in the 2018 dataset was provided by FCAC and was calculated by a third party. The questions used in calculating the index might still be present in the dataset, further aiding the model's prediction. However, we also calculated the well-being score for the post-COVID dataset using the method described in the previous section, and the accuracy was over 95% when we used the same questions to train the model. Thus, we decided to drop these questions and explore other factors that could aid prediction. Although the current post-COVID models may not perform as well as the pre-COVID model, they are still realistic and can generate useful features. In Table 2, we have provided a sample of the top 3 feature importance of each model, listed in descending order. Further discussion of the important features will be presented in section 5.

## 5 DISCUSSION

The discussion is divided into four blocks. The first block gives a general overview of the distribution of the score across the nation and across gender. The second and third block outline the top three features before and after COVID-19, which the study identified as main drivers of FWB in Canada, respectively. The last block presents some of the unexpected and interesting features we found worth laying out, give their importance in policymaking.

### 5.1 Overview of the score

From Figure 6, we observe an unequal distribution of FWB across the nation. The provinces of Nunavut and Northwest Territories present a higher average FWB. This phenomenon may be attributed to several factors. While the responsibilities and lifestyle of these provinces' residents may contribute to their comfort, as reflected by their higher average homeownership rates, it is noteworthy that the sample size in these provinces is comparatively smaller than in other provinces like Alberta and Saskatchewan. The team believes that the limited sample size may have influenced the result to some extent.

We also note from Figures 6 and 7 that the index is also unequally distributed across gender. Males generally appear better off than other genders, particularly before the age of 54. This inequality in FWB distribution is also confirmed at the provincial level (Figure 6), with Northwest Territories topping the list. In other words, a good proportion of the high FWB score observed in the Northwest Territories (see map) is driven by men.

### 5.2 Pre-COVID top features

The above features demonstrate that prior to the COVID-19 pandemic, financial security was primarily determined by the respondents' ability to meet unforeseen expenses and their capacity to enjoy life. Furthermore, home ownership status was a significant determinant, with individuals who own a home having a 50% chance of being financially secure and less than a 2% likelihood of being financially struggling.

### 5.3 Post-COVID top features

Post-COVID, the determinants of financial stability have undergone a shift. It is now determined by whether or not individuals had to rely on credit cards, overdrafts, or borrow money to meet basic expenses like purchasing food or paying monthly bills. In addition, credit score has emerged as a key determinant in predicting FWB status. Moreover, the ability to meet financial commitments and pay bills on time has emerged as a clear indicator of an individual's financial security/struggle.

### 5.4 Interesting features

In our analysis, we discovered unexpected features that contribute to predicting financial security such as the mode of the survey,

**Table 1: Performance summarization, where a. pre-COVID 19 model where we will use the FCAC 2018 dataset. b. post-COVID 19 model with no feature matching. c. post-COVID 19 model with feature matching**

| | **Precision** | | | **Recall** | | | **F1-score** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | a | b | c | a | b | c | a | b | c |
| **Secure** | **0.91** | 0.32 | 0.23 | **0.96** | 0.44 | 0.25 | **0.93** | 0.37 | 0.24 |
| **Somewhat Secure** | **0.92** | 0.73 | 0.70 | **0.88** | 0.76 | 0.72 | **0.90** | 0.74 | 0.71 |
| **Struggling Somewhat** | **0.86** | 0.66 | 0.62 | **0.87** | 0.54 | 0.55 | **0.86** | 0.59 | 0.58 |
| **Struggling a lot** | **0.86** | 0.18 | 0.14 | **0.83** | 0.28 | 0.21 | **0.84** | 0.22 | 0.17 |
| **Macro avg** | **0.98** | 0.47 | 0.42 | **0.89** | 0.50 | 0.43 | **0.89** | 0.48 | 0.43 |
| | a | | | b | | | c | | |
| **Accuracy** | **0.90** | | | 0.71 | | | 0.68 | | |

**Table 2: Top 3 questions that helped in predictions for each model**

| |
|---|
| **a. pre-COVID 19 model** |
| 1. Will you be able to meet an unexpected expense that is equivalent to a month's income? |
| 2. How would you describe your household's/your current financial situation? |
| 3. How does 'Our/my finances allow me to do the things I want and enjoy life' describe you? |
| **b. post-COVID 19 models with no feature matching** |
| 1. In the past 12 months, have you run short of money and had to use a credit card, overdraft, or borrow to buy food or to pay monthly expenses |
| 2. How would you describe your bills and other financial commitments? |
| 3. How would you rate your current credit record? |
| **c. post-COVID 19 models with feature matching** |
| 1. In the past 12 months, have you run short of money and had to use a credit card, overdraft, or borrow to buy food or to pay monthly expenses |
| 2. How would you describe your bills and other financial commitments? |
| 3. Over the past 12 months, have you used any of the following other methods to manage your day-to-day expenses? (Borrow, seek advice, use online lender) |

the marital status, disability status, being a First Nation, and the language you learned as a child. One such feature is the mode of the survey, either online or phone-based, which was found to be an indicator of financial status. Respondents who completed phone-based surveys were 15% less likely to report struggling than those who completed online surveys. This may be attributed to respondents feeling more comfortable sharing honest information about their financial status online rather than with another person over a phone call, as they may fear being judged. This finding highlights the importance of considering the mode of data collection in FWB surveys to obtain more accurate results.

## 6 CONCLUSION

FWB plays a critical role in driving prosperity and livelihood. A high well-being score means a strong capacity to meet financial needs and invest for a better future. On the other hand, a low index is a synonym of financial strains and incapacity to achieve financial objectives. With their effect on the economy as a whole, global shocks like COVID-19 can considerably affect FWB.

Using pre-pandemic and post-pandemic surveys on FWB, this research aimed to measure the drivers of FWB in Canada and investigate the predicting power of machine learning. The results of the study can be summarized as follows:

- COVID-19 contributed to lowering Canadians' FWB
- FWB is disproportionately distributed across the nation and across gender, with economically dynamic provinces and females appearing the most vulnerable.
- The key drivers of FWB include: household financial situation, capacity to meet monthly expenses, saving and credit score.
- The prediction of FWB has a higher level of accuracy when global shocks are not accounted for.
- The low level of accuracy achieved when considering shocks like COVID-19 demonstrates the significant disruption caused

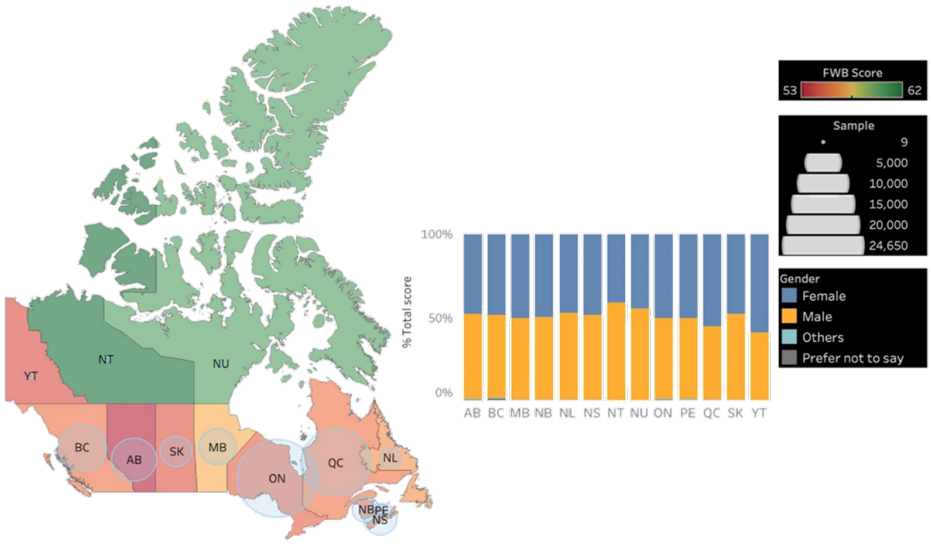


**Figure 6: FWB score along the Canadian provinces, COVID-19 dataset**

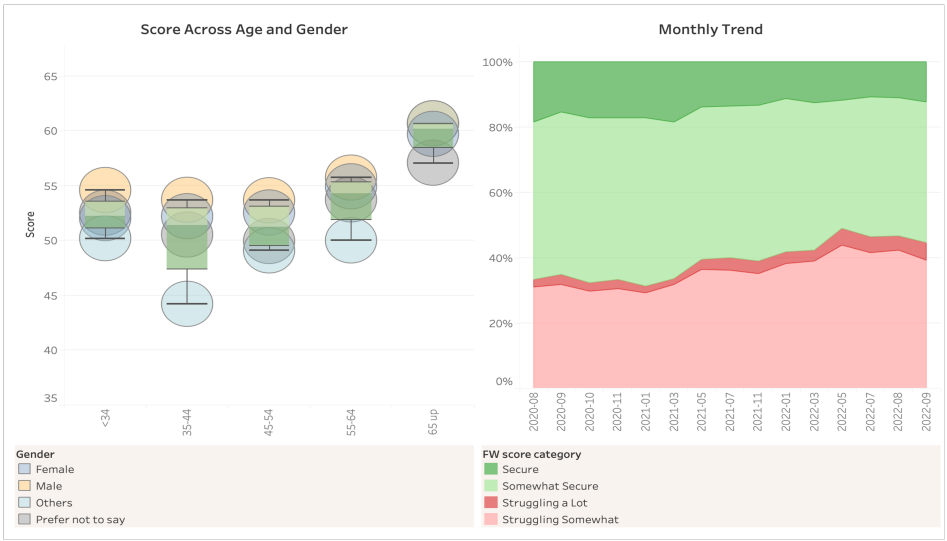


**Figure 7: FWB across gender and time, COVID - 19 dataset**

by the pandemic on a spectrum of sectors including finance, business, job market, health, education etc.

This project advocates for an increasing effort from the government to help Canadians across the provinces better cope with their finances. Addressing this issue will involve creating better jobs, a review of minimum wages, controlling inflation, supporting debt relief including mortgage, further assistance to households in the lowest income deciles. Particular attention should be given to groups that are usually left-behind such as women and single mothers, persons living with disability and indigenous communities.

## ACKNOWLEDGMENT AND DISCLAIMER

We are very grateful to the Financial Consumer Agency of Canada (FCAC) who has contributed all of the data used in the analysis for this project. Any findings, views or opinions expressed in this study are those of the authors and do not necessarily reflect those of the FCAC or the policy or position of any department of the Government of Canada.

## REFERENCES

[1] [n. d.]. Data Spotlight: Financial well-being in America, from 2017 to 2020. https://www.consumerfinance.gov/consumer-tools/educator-tools/financial-well-being-resources/data-spotlight-financial-well-being-in-america-2017-2020/.

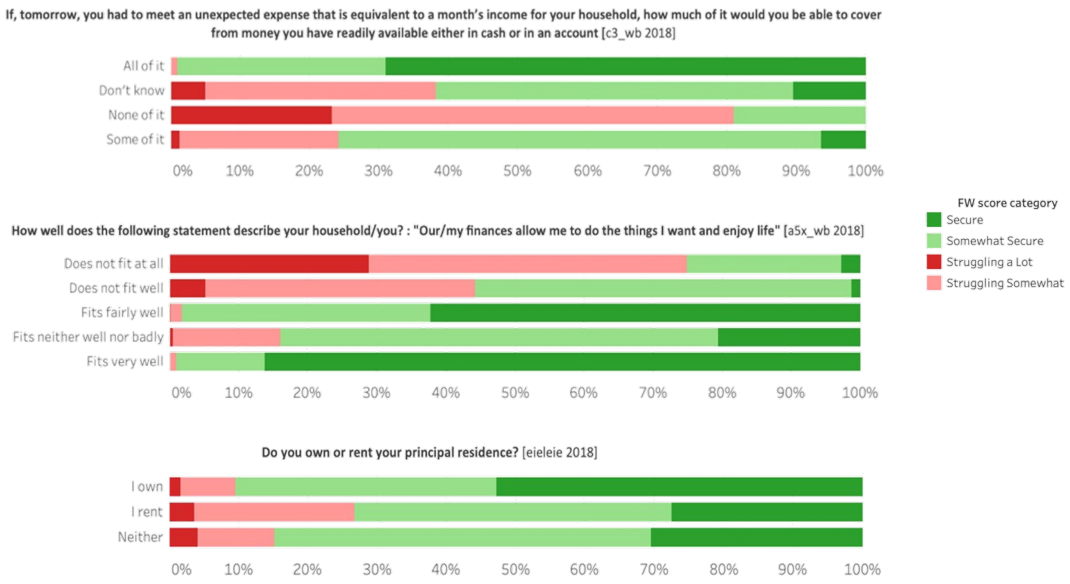


**Figure 8: Pre-COVID top features, 2018 dataset**

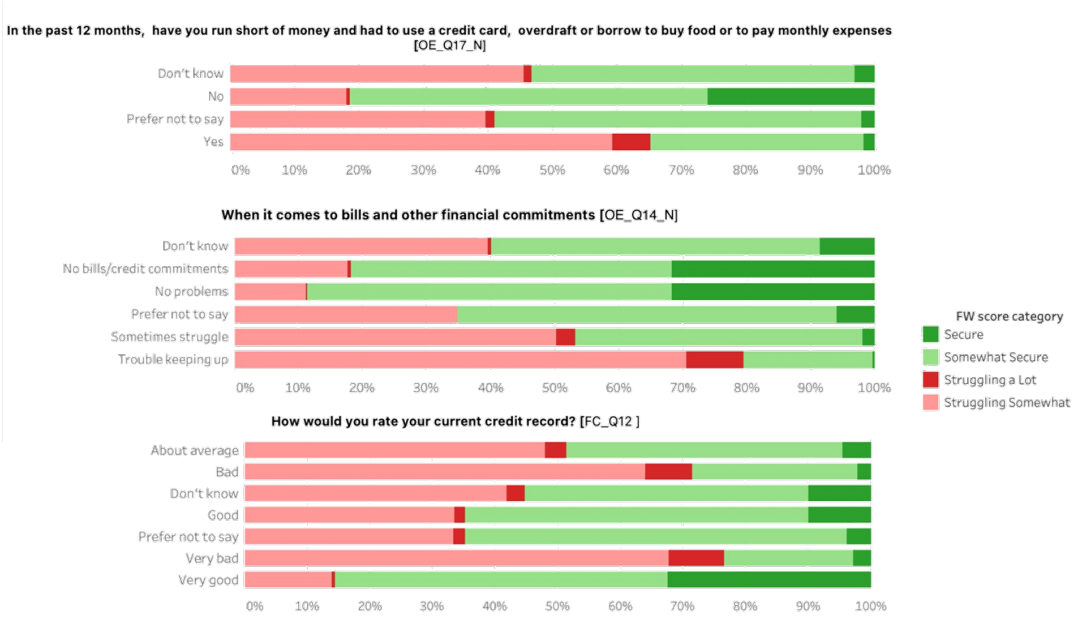


**Figure 9: Post-COVID top features**

Accessed: 2023-04-13.

[2] Jean Bergeron, Luc Ricard, and Jean Perrien. 2003. Les déterminants de la fidélité des clients commerciaux dans l'industrie bancaire canadienne. *Canadian Journal of Administrative Sciences/Revue Canadienne Des Sciences de l'Administration* 20, 2 (2003), 107–120.

[3] Consumer Financial Protection Bureau. 2015. Financial Well-Being: The Goal of Financial Education. Washington, DC: Consumer Financial Protection Bureau.

[4] Consumer Financial Protection Bureau. 2015. Measuring financial well-being: A guide to using the CFPB financial well-being scale. *Washington, DC: Consumer Financial Protection Bureau* (2015).

[5] Angus Campbell, Philip E. Converse, and Willard L. Rodgers. 1976. *The Quality of American Life: Perceptions, Evaluations, and Satisfactions.* Russell Sage Foundation.

[6] Financial Consumer Agency of Canada. 2018. Government of Canada. https://www.canada.ca/en/financial-consumer-agency/programs/research/financial-well-being-survey-results.html

[7] Financial Consumer Agency of Canada. 2022. Government of Canada. https://www.canada.ca/en/financial-consumer-agency/corporate/covid-

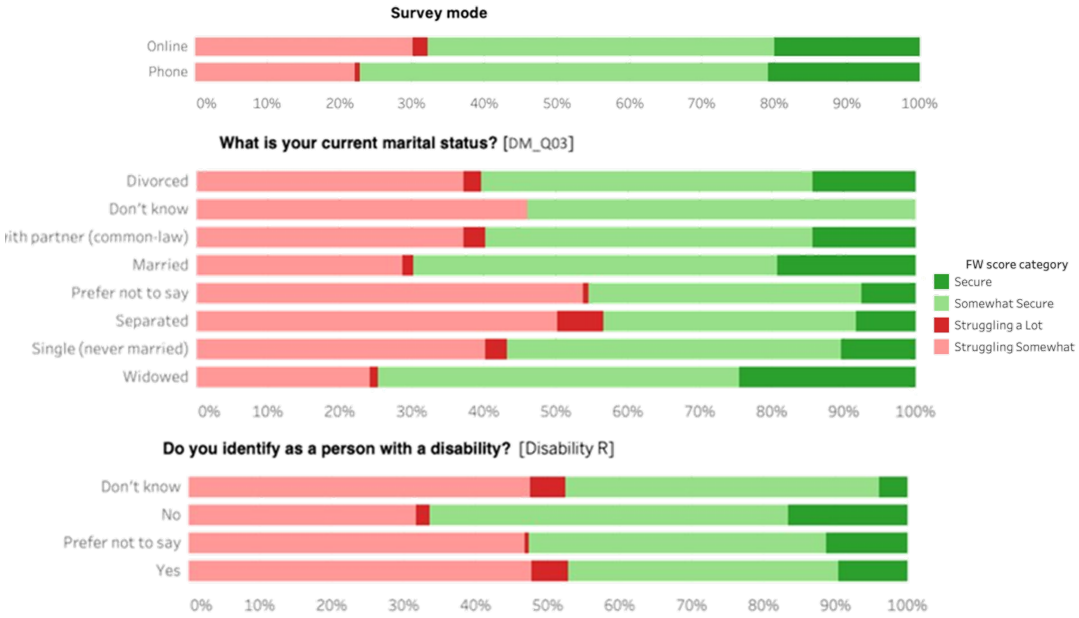


**Figure 10: Interesting feastures**

19/summary-covid-19-surveys.html#toc2

[8] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[9] Elaine Kempson, Andrea Finney, and Christian Poppe. 2017. Financial Well-Being: A Conceptual Model and Preliminary Analysis.

[10] Petro Liashchynskyi and Pavlo Liashchynskyi. 2019. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059* (2019).

[11] Georgios A. Panos and John O.S. Wilson. 2020. Financial literacy and responsible finance in the FinTech era: capabilities and challenges. *The European Journal of Finance* 26, 4-5 (2020), 297–301.

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[13] Ravid Shwartz-Ziv and Amitai Armon. 2021. Tabular Data: Deep Learning Is Not All You Need." arXiv. *arXiv preprint arXiv:2106.03253* (2021).