Housing Market Strategy Research in Canada

Elric Lazaro, Kaiyue Wu, Mengxin Zhao, Zikun Lei

10/19/2020

Code and data supporting this analysis is available at:
https://github.com/ElricL/Housing-Market-Strategy.git

Abstract

While doing business, it is important to determine the target customers. In this study, a logistic regression model is run to explore the relation of one's desirability of owning a house to one's features. It turns out that people's marital status and family income levels are the most significant factors that affect the likeliness of one owning a house along with the person's age. The real estate agents can use this result to effectively find and locate target customers.

Introduction

Business is between dealers and customers. If the potential target customers are found, the whole process would be much more efficient. The same rule applies to real estates. In the housing market, there are two types of customers - one looks to rent, and the other one looks to purchase a house.If real estates agents are able to know whether the customers under certain conditions prefer renting or owning a house, they can target the potential customers and provide better services to increase the chances of getting the trade done.

In this study, the relationship of people's housing preferences to their age, marital status, income, number of children they have, average hours of work and family income(before tax) is investigated, based on the data from General Social Survey with a focus on family, year of 2017.

In this report, we will first explain how the sample data was collected, and specify the explanatory and response variable. Next, we will fit a logistic regression model, and will analyze the results and discuss the model's usage in reality. At the end, we will point out some limitations of our model and provide a plan for the next steps.

The next part focuses on how data are collected and cleaned. Within the same part, the predictor and response variables are specified. After that, a logistic regression model is fitted, and there is an analysis of the results and the usage of the model in reality are discussed. The last part of this report involves limitations of the model and improvements that could be made.

Data

General Social Survey on Family, year of 2017 was selected among the Canadian General Social Survey program. It is used to gather data, thus can be used to monitor changes in Canada and provide information on social policy issues. Among all the survey. Family is closely related to people's living conditions and therefore, the surveys on this topic are informative. Based on further inspection with the 2017 survey, we have discovered that it was given to people aged 15 or older.

The population we seek to study is the whole population of Canada with the frame being those that took that were eligible to fill out a GSS Survey, The sample we have in this study is all those that responded to the 2017 survey which are people aged 15 or older.

From the 2017 dataset we have selected own_rent, income_family, total_children, age, average_hours_worked, and marital_status. Based on the label dictionary obtained along with the data, each variables represent the following:

- **own_rent - "Dwelling - Owned or rented"**
- age - "Age of respondent with decimal at time of the survey interview"
- Marital_status - "Marital status of the respondent"
- Total_children - "Total number of children reported by respondent"
- average_hours_worked - "Average number of hours worked per week"
- income_family - "Family income - Total (before tax)"

With these variables we believe, they would provide useful insights with our goal of finding potential target customers in the housing market. Specifically we would want to see the chances of one to own a place (own_rent) based on the answers given in the other variables. Discovering correlation or independence can both provide useful insights to our research. Along with the data, we have also taken into account the total population of Canada subtracted by the total population of people aged 14 or younger. These numbers will help the model we will be using to analyze the effects of the characteristics chosen on own_rent.

The raw data of General Social Survey 2017 (GSS 2017) was obtained from CHASS. which was then processed and cleaned using a R script code written by Rohan Alexander and Sam Caetano from University of Toronto. Further modifications consist of extracting our variables of observation from the cleaned GSS data to a separate table and removing uninterpretable answers. These unusable answers mainly consist of 'NA' and 'Don't know' values. Unfortunately, these examples were abundant, resulting in the sample size decreasing to almost half the original size. Also important to note that they may still be integral but given their ambiguity, there's currently no way to assess them at the moment.

Since our model requires a boolean variable (a variable that outputs true or false) as response and own_rent is categorical, a new variable, 'own_or_not' was created and it is set to be 0 if own_rent in the same row is in the category "Rented, even if no cash rent is paid" else 1. And this variable is treated as the response variable. Examples of survey inputs on our modified data can be seen below:

total_children	own_rent	average_hours_worked	income_family	age	marital_status	own_or_not
1	Owned by you or a member of this household, even if it i...	30.0 to 40.0 hours	$25,000 to $49,999	52.7	Single, never married	1
5	Owned by you or a member of this household, even if it i...	50.1 hours and more	$75,000 to $99,999	51.1	Married	1
0	Rented, even if no cash rent is paid	30.0 to 40.0 hours	$50,000 to $74,999	28.0	Living common-law	0
0	Owned by you or a member of this household, even if it i...	30.0 to 40.0 hours	Less than $25,000	63.8	Single, never married	1
0	Owned by you or a member of this household, even if it i...	0.1 to 29.9 hours	Less than $25,000	15.7	Single, never married	1
0	Owned by you or a member of this household, even if it i...	50.1 hours and more	$25,000 to $49,999	40.3	Single, never married	1

(Table continues below)

Figure 1: The first 6 rows of the modified data. Notice own_or_not is 1 when survee answers own_rent as owned by them or one of the household members and 0 when they're at most renting.

To help visualize the data a bit more, here is a simple plot of the number of people that owns a house and those that don't:

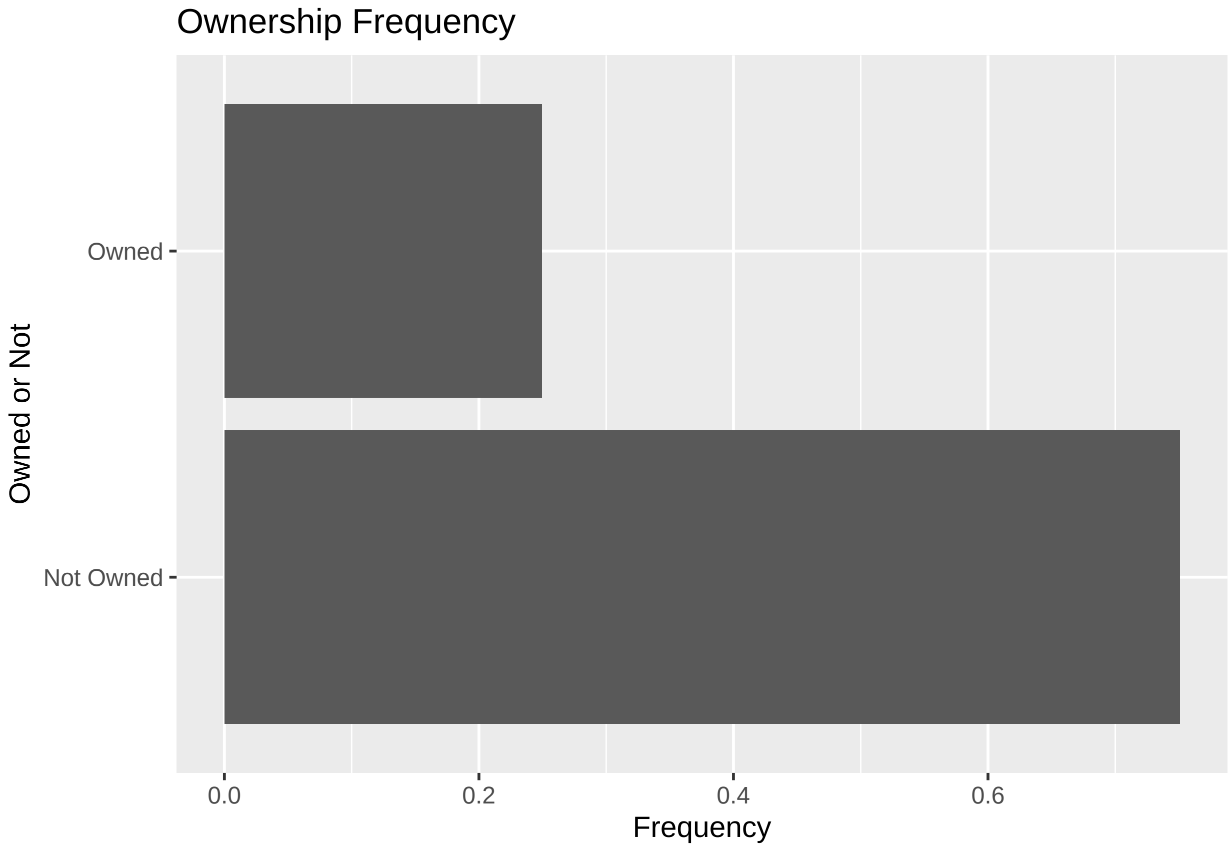

Figure 2: We are interested in finding common characteristics from house owners and non-owners (i.e. what makes a person likely go to each bar), to build a strategy in finding new potential customers)

Model

Model:

Rstudio is used to run the following logistic regression model

$$
\begin{aligned}
log(\frac{(p)}{1-(p)}) = \beta_0 + \beta_1 X_{income\_family:\$125000\ and\ more} + \\
& \beta_1 X_{income\_family:\$100000\ and\ \$124999} \beta_2 X_{income\_family:\$25000\ to\ \$49999} + \\
& \beta_3 X_{income\_family:\$50000\ to\ \$74999} + \beta_4 X_{income\_family:\$75000\ to\ \$99999} + \\
& \beta_5 X_{income\_family:less\ than\ \$25000} + \beta_6 X_{total\_children} + \\
& \beta_7 X_{age} + \beta_8 X_{average\_hours\_worked:0.1\ to\ 29.9\ hours} + \\
& \beta_9 X_{average\_hours\_worked:30.0\ to\ 40.0\ hours} + \beta_{10} X_{average\_hours\_worked:40.1\ to\ 50.0\ hours} + \\
& \beta_{11} X_{average\_hours\_worked:50.1\ hours\ and\ more} + \beta_{12} X_{marital\_status:Living\ common-law} + \\
& \beta_{13} X_{marital\_status:Married} + \beta_{14} X_{marital\_status:Separated} + \\
& \beta_{15} X_{marital\_status:Single,\ never\ married} + \beta_{16} X_{marital\_status:Widowed}
\end{aligned}
$$

Subscripts of X are names of predictor variables or the category if one predictor is categorical. And every β is the slope of the corresponding feature. The value of each β is listed in the table in the result part.

The population size is set to be $36708083 - 1941873 - 2021564 - 1948681 = 32595965$. Where 36708083 is the capital population of Canada in 2017. 1941873, 2021564, 1948681 are populations in the 0~4 years,5~9 years and 10~14 years age groups, respectively. Canadians in those age groups are excluded since every individual was at the age of 15 or above at the moment they took the survey. And so the result cannot represent Canadians in the 0~15 years age group.

The columns income_family, total_children, age, average_hours_worked and marital_status were added into this model as predictor variables and own_or_not is created based on own_rent and be added as the response variable.

The model is mathematically straightforward and the dependent variable, own_or_not, is boolean and numerical. It is mathematically straightforward for two reasons. The first reason is that $f(p) = log(\frac{p}{1-p})$ is strictly increasing on $(0,1)$. Its derivative is $f'(p) = \frac{1}{p(1-p)^2} > 0$ on $p \in (0,1)$ where the probability is usually defined on. And a strictly positive derivative implies that whenever $f(p_1) \geq f(p_2)$, we have $p_1 \geq p_2$. So after sub in two sets of conditions, the one with higher outcome indicates that one under that set of conditions is more likely to own a house. No explicit computation is required to compare the actual probability.The second reason is that, by the property of the logarithm function, $log(\frac{p}{1-p}) \geq 0$ if and only if $\frac{p}{1-p} \geq 1$ if and only if $p \geq \frac{1}{2}$. Thus we can tell if one under a set of conditions is likely to own a house by comparing the value of the dependent variable with 0 given those conditions.

Most Importantly we can also first confirm if there is a relationship between our dependent variable own_or_not and our predictor variables. This can be achieved by obtaining the p-value of each variable in the regression model. The significance of the p-value is that it can tell us that if it has a value less than 0.05 then we can reject the null hypothesis being tested which is if the coefficient is equal to 0 or in other words has no effect.

One limitation of the logistic regression model is that the relation between the dependent variable and features is assumed to be linear. However, a model would still be fitted regardless of the linearity between the dependent variable and features, and a fitted model is inaccurate if the relation is not linear.

Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5647057	0.5288667	1.0677655	0.2856462
as.factor(income_family)$125,000 and more	0.4433885	0.0961359	4.6121012	0.0000040
as.factor(income_family)$25,000 to $49,999	-1.6684636	0.0907608	-18.3830924	0.0000000
as.factor(income_family)$50,000 to $74,999	-1.1041029	0.0901314	-12.2499314	0.0000000
as.factor(income_family)$75,000 to $99,999	-0.5170765	0.0944138	-5.4767039	0.0000000
as.factor(income_family)Less than $25,000	-1.9120066	0.1053783	-18.1442146	0.0000000
total_children	0.0741129	0.0217680	3.4046672	0.0006645
age	0.0266305	0.0019174	13.8890598	0.0000000
as.factor(average_hours_worked)0.1 to 29.9 hours	-0.1981256	0.5107100	-0.3879414	0.6980657
as.factor(average_hours_worked)30.0 to 40.0 hours	-0.4052607	0.5091560	-0.7959461	0.4260779
as.factor(average_hours_worked)40.1 to 50.0 hours	-0.0903597	0.5130483	-0.1761233	0.8601999
as.factor(average_hours_worked)50.1 hours and more	-0.0272966	0.5161192	-0.0528882	0.9578218
as.factor(marital_status)Living common-law	0.2051410	0.0980909	2.0913358	0.0365174
as.factor(marital_status)Married	0.7115567	0.0855797	8.3145536	0.0000000
as.factor(marital_status)Separated	0.0950970	0.1327115	0.7165693	0.4736528
as.factor(marital_status)Single, never married	0.0955210	0.0923321	1.0345373	0.3009043
as.factor(marital_status)Widowed	0.5978007	0.1668188	3.5835325	0.0003402

Figure 2: Key statistical summaries of the logistic regression model.

Before observing the relationships between our chosen predictor variables with whether they own a place or not, we must first confirm the existence of the relationship first. Upon close inspection of the table above that summarize the results of our model, we can see that the following variables have p-values greater than 0.05: all categories for average_hours_worked, and if their marital_status is 'Separated' or 'Single'. This result tells us that we cannot reject the null hypothesis such that there is no effect. Lastly, for each categorical variable, one category is omitted. The slope of every omitted category is 0. And the slope of other categories of the same variable indicate how they affect the response compared with the missing category.

To see which Canadian is unlikely or likely to own a place we can check the estimate column. This column provides the coefficients for our logistic regression function specified in the beginning of the model section of this report. Specifically we can substitute the corresponding variable's β symbol. When filling in the values we can observe that the probability of house ownership will decrease with factors that have negative estimates. Therefore we can conclude the following characteristics will decrease the chance of a person owning a place: having Family income of $\$0\ to\ \$99,999$. The rest of the independent variables have positive estimates and as such we can conclude that people who fall in such categories have a likelier outcome of owning a place. For instance, we can see in the plot below that the total house owners with family income of $125,000 far exceeds the rest. This matches our finding that the factor of income_family of $\$125,000\ and\ more$ having a positive estimated value in comparison to the rest which have negative values.

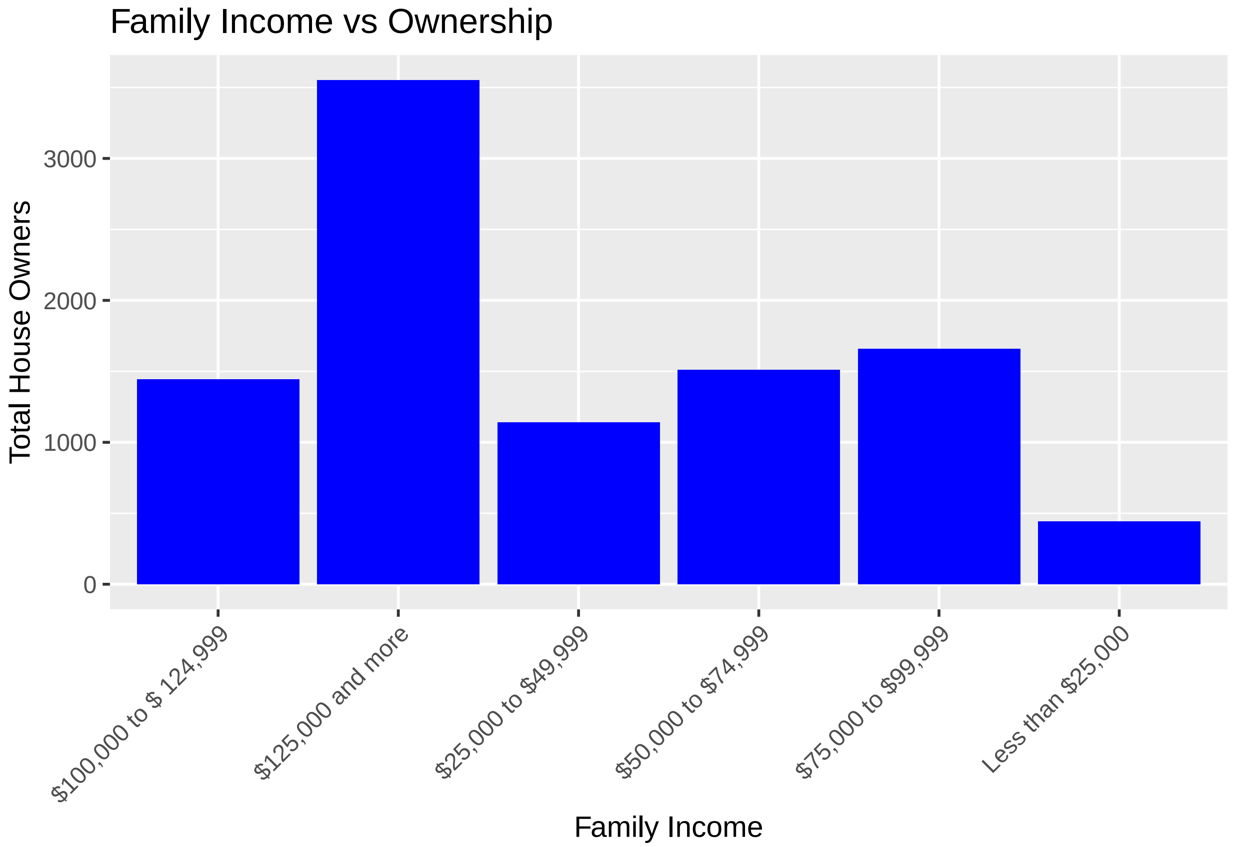

Figure 3: Bar graph of number of house owners based on family income.

Discussion

After applying the logistic regression model on the data which were selected from the General Social Survey on Family, year of 2017, the relationship between one's desirability in owning a house and one's features is investigated. The estimates and p-values for each coefficient are shown in figure 2.

Among all the features, being married, being widowed, and having $125,000$ or more annual income appear to be the most positively influential factors given their high estimated values. Married couples do have a higher chance in purchasing a house, since the financial burden would be lighter for two people, and they need to prepare for life with kids later on. The widowed men and women used to be in the same situation. It also makes sense for people with higher income to own a house.

In contrast, having annual income less than $25,000 and from $25,000 to $49,999 are the most negatively influential factors. For the people who have low income, they probably cannot afford to purchase a house. Renting becomes their only option.

Age's slope is 0.0266. It indicates a positive relation with one's desirability of owning a house. The reason might be that, as people grow older, they usually have higher income. Also, the slope is small since it is a long and slow process for one's condition to change.

The marital status and income appear to be the most influential factors of the model, both positively and negatively. Since the relationship between one's desirability of owning a house and one's features does exist, this model can be used by real estates and other housing agents to faster locate the potential target customers.

Weaknesses

Data is presumed to be real and accurate at the moment when the data was collected. However, there's a possibility that some of the data may be inaccurate and exaggerate due to personal bias. Also, since the data we chose are from 2017, data may be outdated and results may vary. There's a chance that results can not be repeated and reflect well in 2020 due to the large outbreak of COVID-19.

Further details of secondary data is not available. This limits what we can find out. We will have to do further research and surveys to determine potential lurking variables. The choice of owning a place or not may not translate well in this data to certain edge-case scenarios. For instance, a person who would want or willing to have their own place may not have the choice as they are under a limited work contract that forces them to travel a lot. Such a person can still be labeled as a 'potential customer' or at least 'future potential customer' but given their current situation, they would not be in this study. Unfortunately the current dataset does not provide enough information for these edge-cases. Thus we must take the results we have gained from this analysis as more of a rough estimate than a conclusion.

Finite population correction is used to build the model. The weight of different age groups in Canada are not evenly distributed. But in data_design the weight is set to be equal for each row in the data set. This can have some impact on our model's standard error. In simpler terms accuracy of our data could potentially be negatively impacted as a result.

Rows that have "Don't know" as average work hours or own rent are dropped since it is hard to interpret. However, doing so filters out some features. There are many occupations, including, but not limited to, freelancers, seasonal work and part time jobs that usually do not require fixed working hours. And people doing those jobs might answer "Don't know". As a result, such transformation could influence some bias to the data. Similar problem arises when rows that have "Don't know" as family net income are dropped.

Next Steps

We notice that the p-value of the intercept is pretty large. So a test comparing the mean of own_or_not with 0 is necessary.

As previously mentioned there's the possibility of lurking variables that we have yet to observe. Therefore a similar analysis will need to be conducted on various other datasets outside of GSS.

The data and the analysis performed is a summarization of Canada as a whole. While this is an important start, the housing market varies in different geological areas around Canada. To look into this we would need to perform a different model in the spirit of stratified sampling where we would need to analyze the likelihood of place ownership in different regions.

Another step we would need to take is to look into the edge cases mentioned in the weaknesses section.

Due to the complexity and the fact that we currently do not have a grasp of the amount of common cases there are various surveys will need to be performed to have a better understanding. Depending on the results obtained, we could perform a similar model.

As a result of the current pandemic in Canada and the whole world and its effect in today's climate, this study would need to be repeated again in an updated General Social Survey on Family Data in 2020.

References:

1. General Social Survey: An Overview, 2019. (2019, February 20). Retrieved from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

2. Tidyr. (n.d.). Retrieved from https://www.rdocumentation.org/packages/tidyr/versions/0.8.3/topics/drop_na

3. kable. (n.d.) Retrieved from https://www.rdocumentation.org/packages/knitr/versions/1.30/topics/kable

4. kableExtra. (n.d.) Retrieved from https://www.rdocumentation.org/packages/kableExtra/versions/1.2.1

5. Alexander, R and Caetano, S. (2020, Oct 7). GSS.cleaning. Retrieved from U of T Quercus

6. Government of Canada, S. C. (2018, March 27). Canada at a Glance 2018 Population. Retrieved from https://www150.statcan.gc.ca/n1/pub/12-581-x/2018000/pop-eng.htm

7. Government of Canada, Statistics Canada. (2020, September 29). Population estimates on July 1st, by age and sex. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501