Analysis_of_Investment_Outcome_Predictor_Report

Nikko Dumrique, Mahdi Heydar, Harry Zhang, Ahmed Rizk

Contents

Summary 1
Introduction 1
Research Question 2
Dataset Description 2

Premliminary Analysis **3**
Methods and Results 5
Discussion of Overall Results: 10
References: 11

Summary

The goal of this analysis is to investigate the economic family's investment income, where the true class indicates money was made on investments, and the false class indicates breaking even or money lost. The report examines its relationship with family size, the economic family's major income earner, the income after tax, and other related variables to determine whether or not its possible to predict it using those variables or a subset of them. The data is obtained from Statistics Canada's 2017 Canadian Income Survey [1].

Introduction

It is estimated that roughly 39% of Canadians invest in stocks, putting Canada in 6th place among 16 countries for share of retail investors [2]. As every generation makes their first investments earlier than their predecessors [3], the importance of gauging and predicting the success of said investments arises.

Families of most wealth levels and sizes have been steadily placing an emphasis on maintaining passive income and generational wealth, through a variety of ways that range from real estate ownership to family offices [4]. In this analysis, we are particularly interested in the family size and its relationship to the investment income.

The economic family is defined by Statistics Canada as "a group of two or more persons who live in the same dwelling and are related to each other by blood, marriage, common-law union, adoption or a foster relationship" [5].

We will examine the investment income of the economic family with the aim of identifying whether or not we can predict it using the economic family size and the major income earner.

There are other variables of interest that may also be useful in the context of the predictive analysis. We hypothesize that the average number of weekly working hours for the major source of income, their highest level of education and their income after tax may affect the economic family's investment income. Furthermore, the economic family's income after tax may be a strong candidate as a predictor variable, although there are some reservations due to it possibly being too collinear with the investment income.

Research Question

- How do economic family size and the major income earner in a family influence familial investment income outcome, where the true class indicates money was made on investments, and the false class indicates breaking even or money lost.
  - It should be noted that the question initially aimed to predict the numerical value of the investment, but the accuracy obtained showed a regression problem to be unsuitable (see report), so the problem was altered to a classification problem.

Dataset Description

Canadian Income Survey (CIS) is a cross-sectional survey sponsored both by the Government of Canada and Statistics Canada. The purpose of this survey is to collect information from all Canadian citizens and households. However, around 2% of the residents in reserve aboriginal settlements or small population, extremely remote areas are not included in this survey. The data was collected with several different characteristics in mind including labour market activity, school attendance, disability, support payments, child care expenses, inter-household transfers. It also combines some information from the Labour Force Survey (LFS), such as the information about the education level and geography. This data set is available to all organizations, different levels of the government, and individuals. Different government levels could use this dataset to make economic well policies to all Canadians.

Description of Relevant Variables

The original dataset contains 194 variables. However, our analysis is only concerned with 8, narrowed down using the aforementioned studies and assumptions. The description contains only those specified variables.

Description of the Relevant Variables:

- EFSIZE: The number of economic family members: It is numeric, discrete and has a range of 1-7.
- USHRWK: The average weekly working hours in the year. It is a numeric variable.
- ATINC: The total annual income before tax for each sampling unit. It is numeric, continuous, and has a range of -70395 to 825710.
- HLEV2G: The highest education level for each sampling unit. It is discrete and categorical, but represented with a number code.
  - 1: Less than high school graduation
  - 2: Graduated high school or partial postsecondary education
  - 3: Non-university postsecondary certificate or diploma
  - 4: University degree or certificate
  - 6: Valid skip
  - 9: Not stated
- EFINVA: The total investment income. It is a numeric and continuous variable. It also includes the net partnership income and net rental income.
- EFMJIE: A binary variable, representing whether or not the individual is the Major income earner of the economic family.
- EFATINC: The total annual income after tax for each economic family. It is numeric, with a range of -39850 to 1128860.
- EFMJSI: A discrete, categorical variable, which represents the major source of income for each economic family. The levels are represented using a number code.
  - 01: No income
  - 02: Wages and salaries
  - 03: Self-employment income

Table 1: The CIS Dataset (first 5)

YEAR	PUMFID	PERSONID	FWEIGHT	PROV	USZGAP	MBMREGP	AGEGP	SEX	MARST	CMP
2017	2129	212901	NaN	24	8	18	4	2	6	
2017	2129	212902	NaN	24	8	18	4	1	6	
2017	2129	212903	NaN	24	8	18	10	2	2	
2017	2129	212904	NaN	24	8	18	10	1	2	
2017	2130	213001	NaN	12	5	7	16	1	1	

  - 04: Government transfers
  - 05: Investment income
  - 06: Retirement pensions
  - 07: Other income

Premliminary Analysis

Loading Data

Cleaning and Wrangling

The data is reduced to some specific features and targets of interest before EDA, using the studies around the research question to narrow them down from 194 features to the most relevant 8.

Variable Data types and Modifications

- All variables have a numeric type, but not all of them are quantitative.
- EFMJIE, EFMJSI, and HLEV2G are categorical variables with levels represented as discrete numeric values.
- Although the USHRWK column is relevant to our analysis, the dataset has 0 zero valid entries. This is due to all values being NaN. This was also expressed in the dataset guide provided by Statistics Canada []. The remaining columns have a complete count of valid entries. (all non-null counts at 92292 which is in line with the number of total observations in the data) including the valid skip.
- The ATINC column contains the value 99999996 for observations that were skipped for the information for valid reasons. Therefore, these observations should be removed.

Comments on the figure

- Most observations have 2 economic family members,
- After tax income for both individuals and the economic family shows a distribution with a slight skew to the left.
- Investment income for the economic family shows a similar distribution to that of the after tax income, which could be a derivative of their potential multicollinearity (Investment income is a part of the total income before tax).
- Most economic families have salaries and wages as their major source of income. As this mode is relatively very large, it may indicate a lack of suitability as a predictor variable for the investment income if a large majority of the surveyed have the same source of income.
- Upon further examination, the education level and the after tax income for individuals will not be used in the further analysis as they are variables relating to the individual and not the economic family as a whole. The lack of interesting trends in the graphs make it so that there is little sense in including them as predictors.

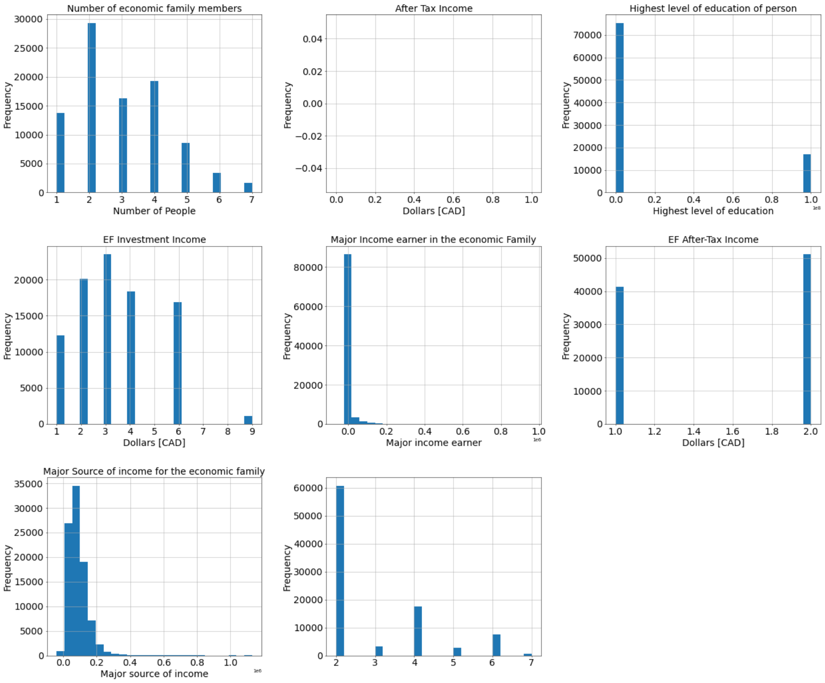


Figure 1: Frequency Distributions of the chosen variables

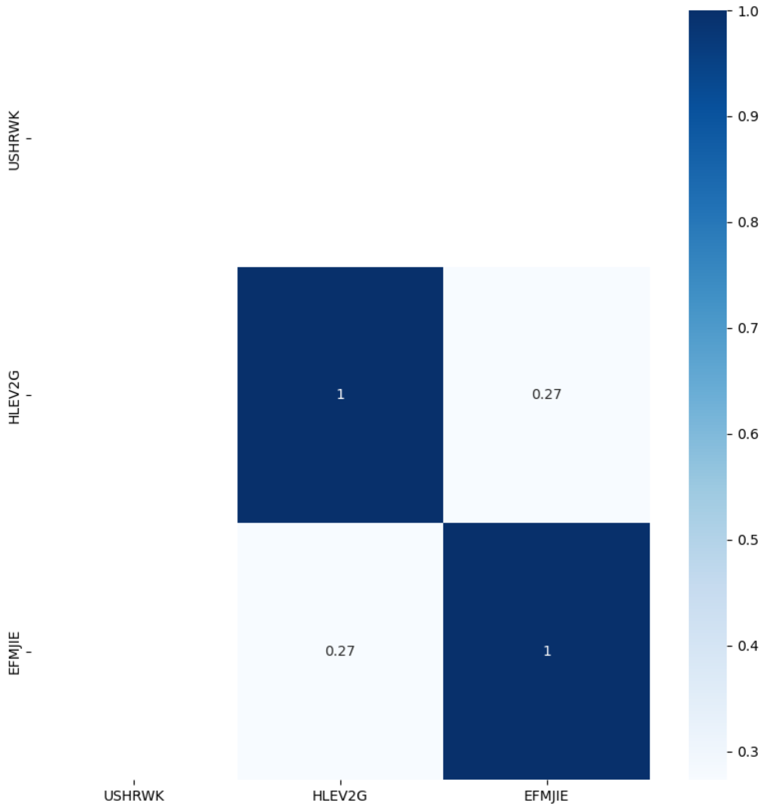


Figure 2: Matrix of correlations between various features

Note: correlations between categorical features should be ignored as these are invalid

Comments on the figure

- ATINC shows a moderate to strong correlation to EFATINC, which can be explained by them representing the same quantity but the former relates to the individual as opposed to the latter's economic family.
- EFINVA shows a moderate to weak correlation with EFATINC. This is unexpected as we did not foresee the tax removal from 'income before tax' affecting the collinearity of the 2 incomes by this degree. This could be an area for potential analysis in a future study.

Methods and Results

Data Splitting and Preprocessing

We decided to remove EFATINC and ATINC from the analysis. The former suffers from a weak correlation with our response variable, while the latter is not related enough in the context of our study.

The ratio of the train/test split was chosen as 70/30 to achieve the best balance between model accuracy and testing accuracy. In order to make sure that we can preprocess our training data, we have to separate it into two parts each representing the X variable and the Y variable. We also repeat process for the test data.

We build a transformer to do ordinal encoding. As for better machine readability, we need to make our variables start from 0 instead of 1.

The First Model: Ridge Regression

Hyperparameter Optimization We use cross validation to improve the prediction accuracy. The tuning process is to find the best alpha value, which is then used to create the model.

Table 5: Cross validation scores for ridge model We found the cross validation (CV) scores as follows. Then we create a reverse elbow plot to find the best number cross validation folds.

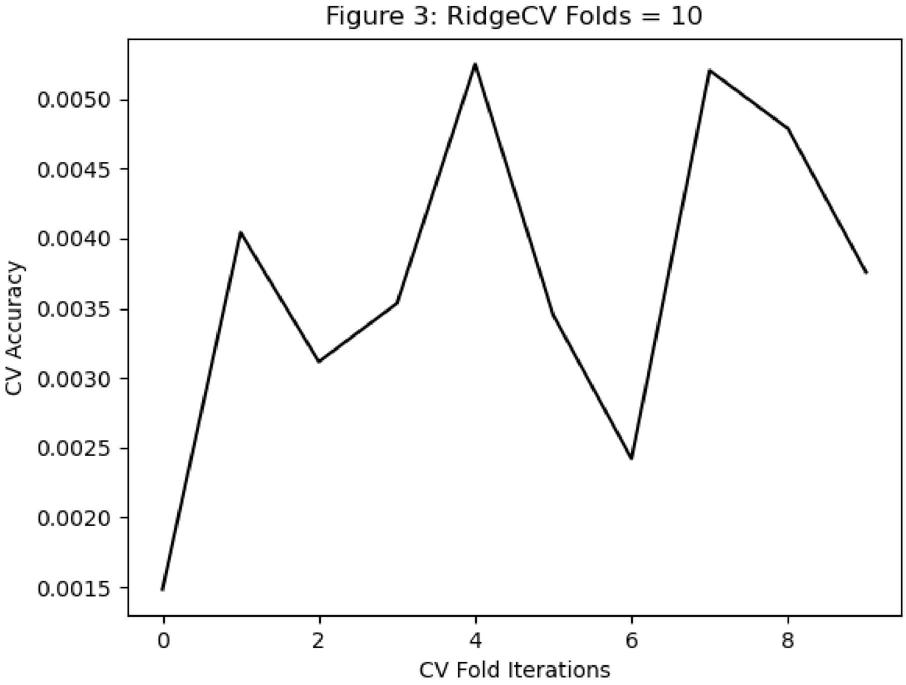


Figure 3: Cross Validation Accuracy vs Fold Iterations

According to 3 we can see that the best CV fold value should be 4, which corresponds to the maximum CV accuracy.

Discussion of the Model Results Although we choose 4 as the best cv fold value, it still shows a extremely low prediction score of approximately 0.004 (as seen in 3) on the testing sets (R^2 value is nearly 0). This shows there is very weak correlation between our predictors and response variables. Meaning, there is little to no relationship between economic family investment income (EFINVA), and the family size (EFSIZE), and the economic family's major income earner (EFMJIE).

Second Model: KNN-Classification

We hypothesize that our model's weakness may be due to the exclusive use of categorical and low range discrete variables. Since our continuous options (EFATINC) and (ATINC) will not be suitable predictors for our response variable in this case, we can change the model to a classification. Building a model that predicts whether or not the economic family made investment profit may be a better method to investigate the relationship given the predictor variables we are working with. Therefore we convert our response variable into the new variable "EFINVA_Made_Money".

The same splitting methodology has been applied here as in the regression model.

Next, we tune the model to find the best value of K neighbors.

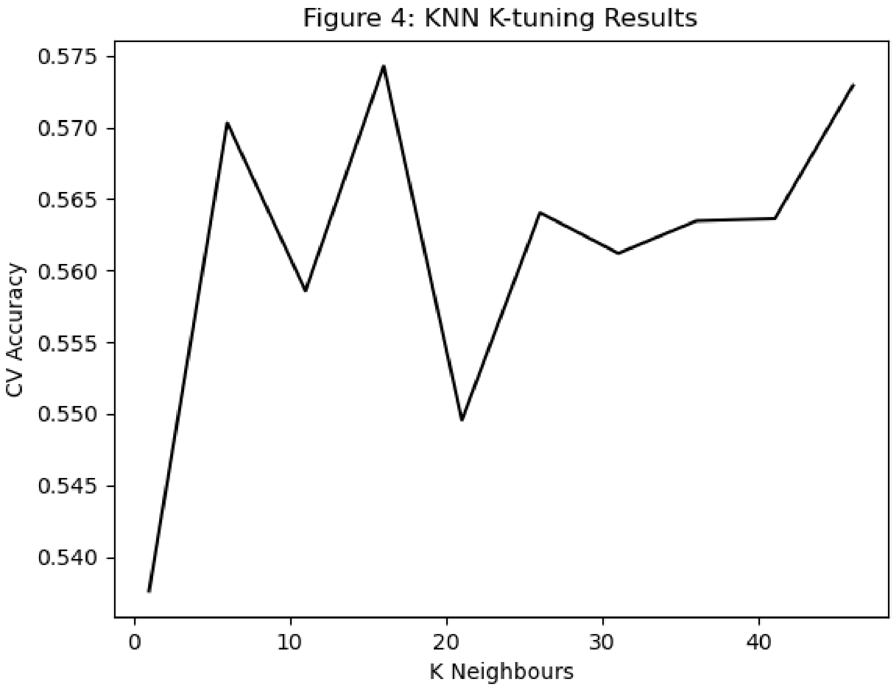


Figure 4: Elbow plot of K-values vs CV accuracy

The reverse elbow plot 4 shows an optimal K value of 26. Even though a higher CV score is obtained at K=16, this is likely just due to chance. K=26 is a better optimal value since the CV score remains relatively stable past K=26. It shows a mean cross validation score of 55.7% which is a very good estimate for our testing score (see below).

Testing Score The model shows an accuracy of `python pickle.load("../result/final_score")` percent on the testing data, which is very slightly better than arbitrary guessing.

Classification Results Visualization

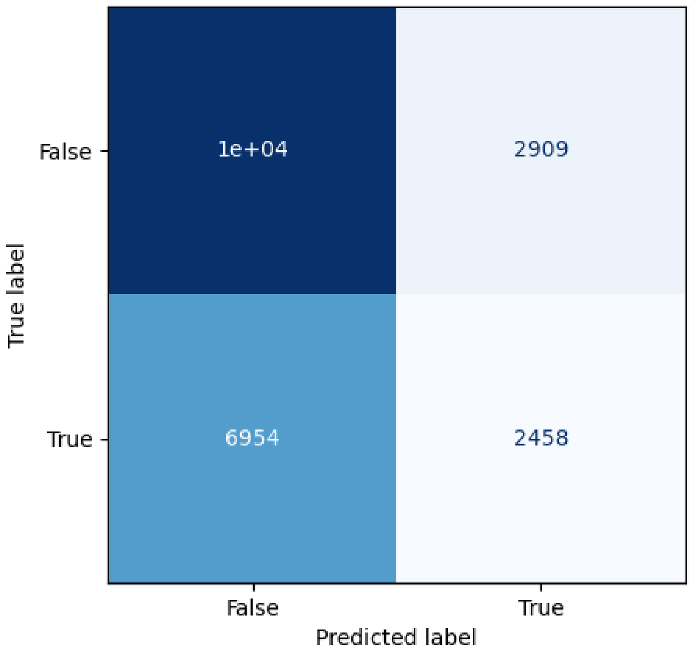
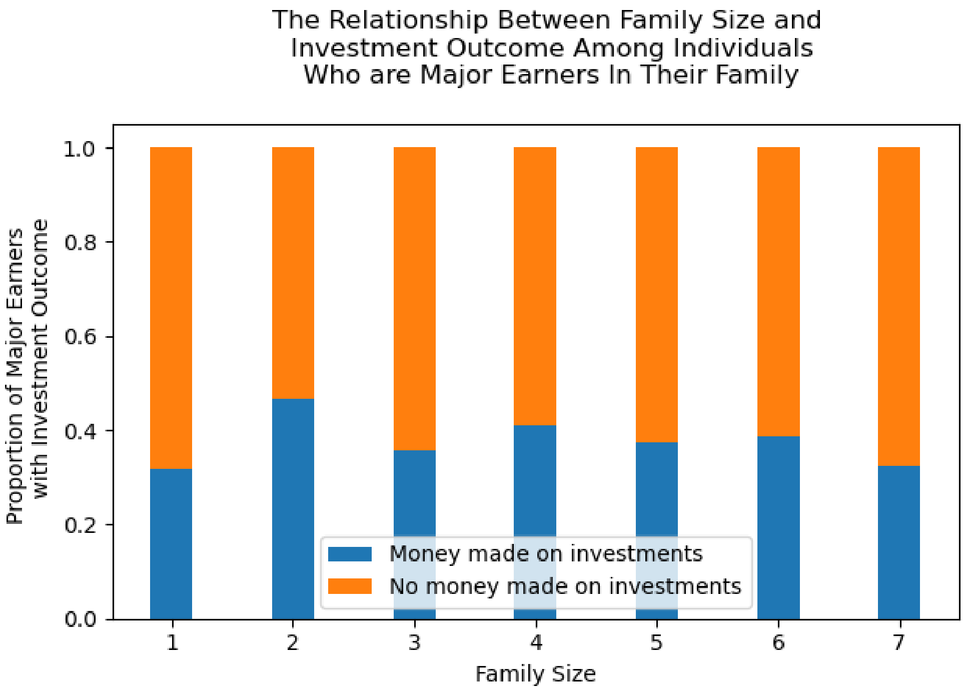
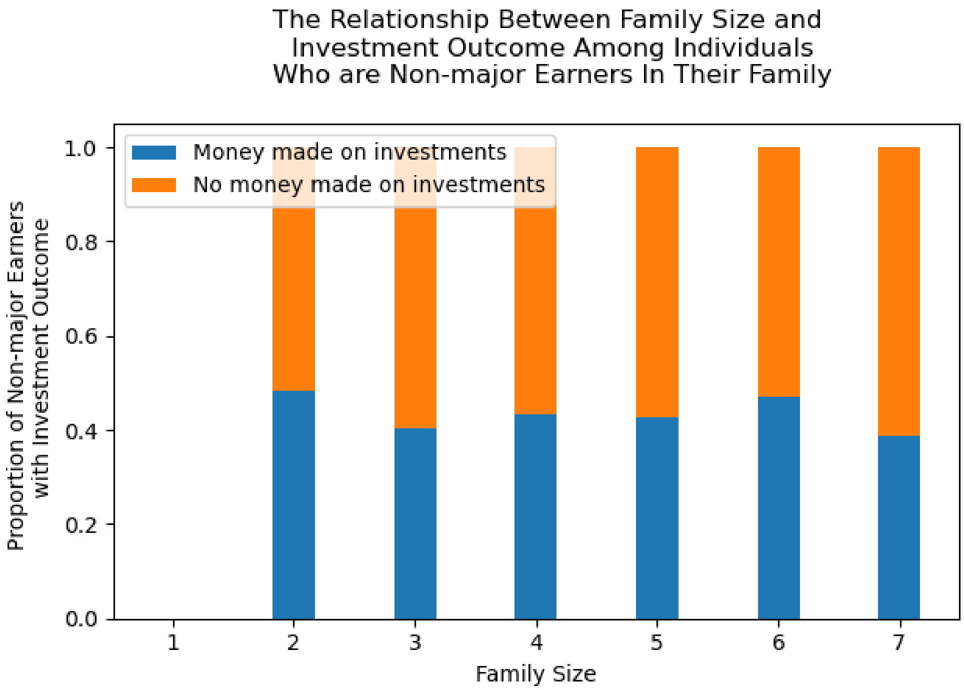
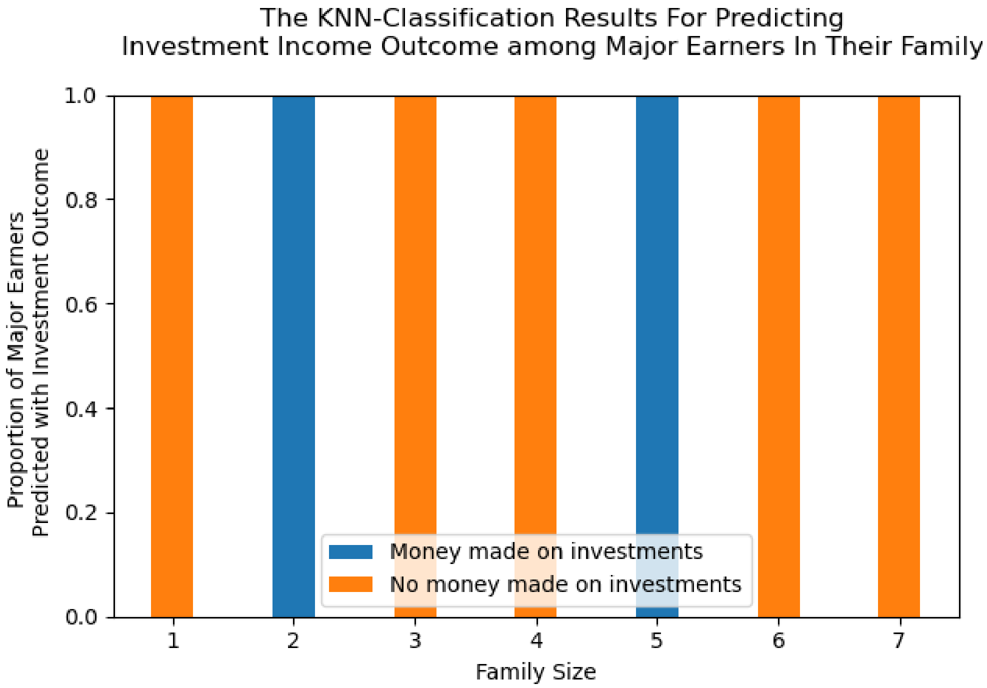


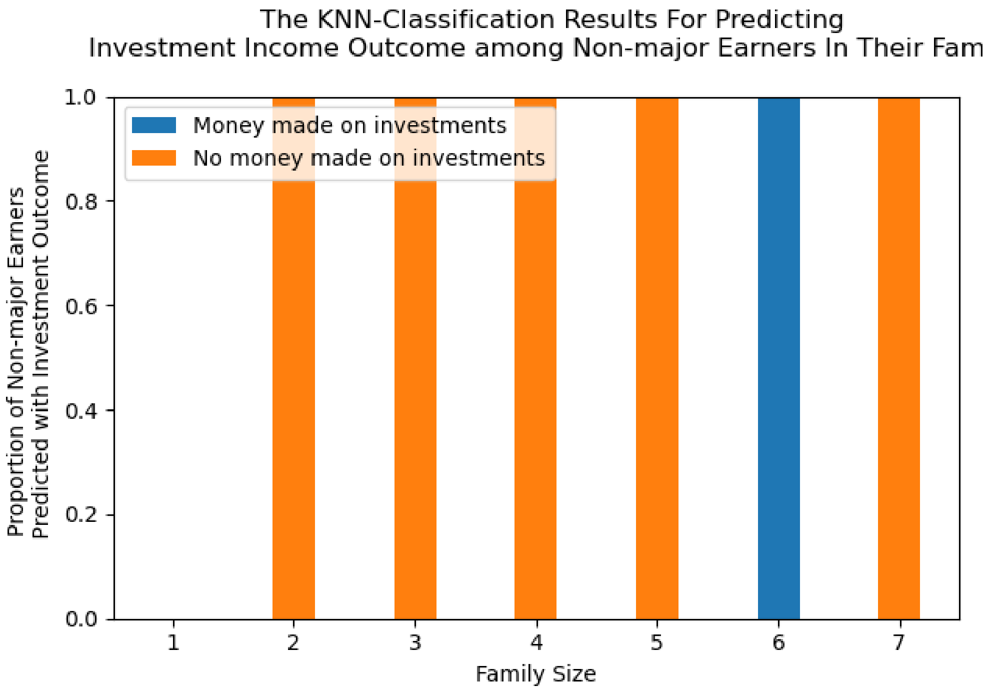
Figure 5: Confusion Matrix



Discussion of Plot Results


The Relationship Between Family Size and
Investment Outcome Among Individuals
Who are Non-major Earners In Their Family
Proportion of Non-major Earners
with Investment Outcome
Money made on investments
No money made on investments
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
Family Size


The KNN-Classification Results For Predicting
Investment Income Outcome among Major Earners In Their Family
Proportion of Major Earners
Predicted with Investment Outcome
Money made on investments
No money made on investments
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
Family Size



?? and **??** show the actual results from the test data for investment income outcomes among major earners and non major earners respectively. Both figures show families of size 2, 4 and 6 to have the largest proportions of people making money from investments while families of 7 have the lowest proportions of individuals making money. However, among the non-major income earners, the proportions of individuals making money is generally slightly higher. Nevertheless, The percentages either breaking even or losing money on their investments is larger among a majority of the family sizes in both cases. A key difference is also present among non major earners where families of 1 are non existant. This is logical within this context since a family consisting of one person automatically makes that person the major earner.

?? and **??** show the predicted results from the KNN-classification algorithm deployed on the test data for major earners and non major earners respectively. The results are not extremely surprising given the ~57% accuracy obtained. The KNN-classification model seems to be unable to distinguish between the different family sizes since it classifies all individuals within a specific family size as the same.

Discussion of Overall Results:

Using two different models, we concluded that the relationship between the Economic Family's Investment Income, size, and major income source is very weak. This was unexpected, as the original assumption of high correlation between the variables was disproven.

The relative improvement in the accuracy of the KNN classification model compared to the Regression model may indicate that the predictor variables are better at predicting whether or not the economic family made profits through investment, than they are at predicting the magnitude of said profits/losses.

The design decision to eliminate the Economic Family's Income After Tax from the list of predictor variables, due to low correlation, may have negatively affected the accuracy of both models. Prior to starting the analysis, it was assumed that it would be too collinear with the response variable. However, the correlation coefficients showed this not to be the case. A possible explanation is that the economic family's income before tax is what suffers from this multicollinearity with the investment income, due to the latter being one of the many contributing incomes to the sum that makes it up. The removal of tax may offset this multicollinearity to a strong enough degree that it would make for better predictive models, and the degree of this change may be an interesting topic for a future study.

A further point of discussion is the age of the dataset, as the data was collected prior to the cryptocurrency boom in 2017. Some of the data regarding the relationships surrounding the investment income (direct or indirect) may be outdated.

In the future, Data Scientists studying investment income of Canadian Families may refrain from dedicating significant resources to using the family's size and major source of income as possible predictors, due to the lack of evidence for a relationship between them exemplified through the analysis.

Additionally, it may encourage more studies surrounding investment income of individuals as opposed to that of families, which may lead to different results due to variables such as the investor's highest education level and sex being more relevant. It may also better fit the investor's status as the major income earner of the economic family, as its lack of relationship with the family's investment income may indicate more suitability in a study about the individual rather than the family

References: