

Retirement Income Adequacy and Workplace Pensions

CKME136 P19 01
Data Analytics: Capstone Course
Melissa Cortina 500928550

CONTENTS

ABSTRACT.....	3
LITERATURE REVIEW	4
DATASET.....	6
APPROACHES.....	10
Technique 1: Analysis of Occupational Type and Industry Type	10
Technique 2: Classification – J48 Decision Tree, Naïve Bayes, K-Nearest Neighbours.....	13
Technique 3: Association Rules – Apriori Algorithm.....	14
RESULTS	15
Technique 1.....	15
Technique 2.....	15
Technique 3.....	15
CONCLUSION.....	16
REFERENCES.....	17

ABSTRACT

The general rule of thumb is that you must have approximately 70% of your salary in your last year of employment for a sustainable retirement income. That income is usually made up of three sources: government-sourced pensions, annuities from personal savings, and your workplace pension.

From the angle of a pension plan, one method to maintain a fully funded status is ensuring that the member enrollments into the pension plan remain constant or increase. Therefore, a pension plan could seek out employers who do not currently offer a workplace pension to their employees. In turn, an employer can assist their employees in achieving the balanced retiree income required upon retirement.

Using data from Statistics Canada, the breakdown of income will be determined at retirement and compared to an individual's last year of employment for specific types of occupations. If that income is less than 70% and there is no sign of an employment pension plan, then their income at retirement is unbalanced. A pension plan could target employers who are part of that particular workforce and propose the offering of a workplace pension to their employees. The data will also be analyzed to determine the common characteristics of those individuals who do have a workplace pension and apply the model when a pension plan is considered enrolling individuals.

LITERATURE REVIEW

The structure of Canada's Retirement Income System (RIS) has evolved from the time of the enactment of the *Old Age Pensions Act* in 1927 until now. However, one thing has remained constant: the need for multiple sources of income during retirement to achieve a replacement ratio that will support a sustainable standard of living during retirement.

The various sources of retirement income can be summarized into three pillars. The first pillar contains Old Age Security (OAS) and Guaranteed Income Supplement (GIS), two programs financed by the Government of Ontario. The second pillar is made up of the mandatory pension programs, one of which is the Canada Pension Plan (CPP). If applicable, the workplace pension plan (WPP) would be part of this pillar. Lastly, the third pillar contains voluntary pension savings programs, such as, personal or group registered retirement savings plans (RSPs) and tax-free savings accounts (TFSA's). (Baldwin, 2009)

Although well-structured, the RIS does not work well for all ranges of income earners. The RIS works differently if you're a high-income earner versus a low-income earner, due to the claw backs in place once you earn a certain threshold of taxable income. Therefore, a part of one of the "retirement pillars" no longer applies as one of your many sources of retirement income that was once incorporated in your planning. The RIS is a puzzle and the puzzle changes depending on an individual's level of income. (MacQueen, 2019)

Approximately one half of the replacement ratio, that is, the percentage of a worker's income that they will earn upon retirement, can be attributed to the public components of the RIS, that is, OAS, GIS, and CPP. Observing the ranges of income earners by using Statistics Canada's Survey of Financial Security (SFS), at higher levels of income, the amount needed for the remaining replacement income increases quite quickly. (Baldwin, 2019) This supports the need for WPPs to assist in achieving an appropriate replacement ratio.

Opposing views suggest that the Canadian RIS is working well and there is no crisis. The reason it was perceived that we were facing a retirement crisis was due to the fact that the sources of future retirement income were being underreported and understated. Although these conclusions were made in 2013, the paper also concludes that the private-sector RPP coverage from WPPs has been declining for decades. (Vettese, 2013).

While the RIS is based on the 3-pillar model, some argue that analysis on the RIS has been so focused on this model, that it has created a false sense of the retirement climate. The 3-pillar model fails to address two other major pillars, which are the assets held outside the pension vehicles, such as, home equity and non-tax preferred accounts, and the undocumented network of family and friends that lend financial support to retirees. By ignoring these factors, it has been perceived that Canada faces a major retirement crisis. (Cross, 2014)

When analyzing all sources of retirement income that are part of the 3-pillar system and outside the traditional model, when an individual does not have a WPP, they will retire with totally inadequate retirement savings. This supports the need for having a WPP. Half of Canadian couples between 55 and 64 have no employer pension between them. (McCarthy, 2016)

More recent studies show that the risk of a declining standard of living during retirement is largely concentrated among the middle to upper income earners, particularly those who are not participating in a WPP. In Baldwin's commentary in 2016, he compares five different studies and their methods to assess the future retirement climate. All arrive at the same conclusion, that the future elderly will experience a decline in their standard of living. (Baldwin, 2016)

The RIS is composed of many sources of income, but another important component is the replacement ratio which addresses how much of these collective sources are required. Just as the 3-pillar model has been in practice for a long time, so has the general rule of thumb that the replacement ratio should be 70% of an individual's income in their last year of employment prior to retirement. A new proposed measurement is the Living Standards Replacement Rate (LSRR) which was introduced because the employment earnings in a single year is not a reliable representation of a worker's standard of living. The goal of the LSRR is to capture a worker's living standards continuity after retirement, by comparing how much money a worker has available to support personal consumption of goods and services before and after retirement. (MacDonald et al., 2016)

Improvements to the RIS have been ongoing, and one improvement in particular is the enhancement of the CPP. A study by Baldwin and Shillington outlines multiple improvements to the RIS and analyzes if the CPP changes will benefit all levels of income. They conclude that it will provide little benefit to low-income earners since it fails to take into account the impact of demographic and labour market changes on the retirement income system. It is problematic because of how the CPP benefits interact with parts of the RIS and the tax system. (Baldwin and Shillington, 2017)

Finally, another possibility to improve the RIS is by modifying one of the existing pillars, as discussed in a research paper through the National Institute of Ageing at Ryerson University. The paper discusses modifying the WPP to something similar to the existing WPPs, except instead of functioning in a registered savings environment, the WPP would function in a tax-free savings environment, similar to a TFSA. Once an individual retires, the pension from the WPP would not count as income after retirement. (MacDonald, 2019)

DATASET

The dataset used for this analysis is the “Survey of Labour and Income Dynamics (SLID), 2011 [Canada]: Person File” from Statistics Canada. It contains survey data that is collected on an annual basis throughout all of Canada. It is a collection of income, labour and family variables on persons in Canada and their families. The samples for SLID are selected from the monthly Labour Force Survey (LFS), and therefore, share the LFS's sample design. The LFS sample is drawn from an area frame and is based on a stratified, multi-stage design that uses probability/random sampling. (Statistics Canada, 2011)

The complete dataset contained 47,705 instances and 147 attributes. The attribute `alfst28` represented a person’s annual labour force status. If the status was “not in labour force”, that is, a code of “3”, then those instances were removed from the dataset since the analysis is with regards to those in the workforce and their employer pension. A review was then conducted on the remaining attributes to determine which attributes would be applicable for the study. At that point, no statistical analysis was conducted to determine which attributes to select; only the description of the attributes from the associated literature for the dataset was reviewed to determine if an attribute was relevant to the topic or not. After this review was completed, it was determined that a total of 26 attributes were relevant to the topic being studied.

The chart below defines the initial 26 selected attributes and their data types: qualitative (nominal or ordinal or interval) or quantitative (discrete or continuous):

Attribute Name	Description	Qualitative or Quantitative	Data Type
<code>ecage26</code>	Person's age as of December 31	Quantitative	Discrete
<code>ecsex99</code>	Sex of respondent	Qualitative	Nominal
<code>marst26</code>	Marital status of person as of December 31	Qualitative	Nominal
<code>mjacg26</code>	Person's major activity at the end of year	Qualitative	Nominal
<code>immst15</code>	Flag if person is an immigrant	Qualitative	Nominal
<code>pvreg25</code>	Province of residence	Qualitative	Nominal
<code>dwtenr25</code>	Ownership of dwelling	Qualitative	Nominal
<code>mortg25</code>	A mortgage on the dwelling	Qualitative	Nominal
<code>multj28</code>	Flag if person holds multiple jobs	Qualitative	Nominal
<code>alhrp28</code>	Total hours paid all jobs during year	Quantitative	Continuous
<code>mtlswk28</code>	Number of months since person last worked	Quantitative	Continuous
<code>yrxfte11</code>	Number of years of experience, full-year fulltime	Quantitative	Continuous
<code>clwkr1</code>	Class of worker for this job	Qualitative	Nominal
<code>prmj1</code>	Flag to indicate if job is permanent	Qualitative	Nominal
<code>flprt1</code>	Flag if job was fulltime	Qualitative	Nominal
<code>nocg2e6</code>	Occupational code	Qualitative	Nominal
<code>imphwe1</code>	Implicit hourly wage for this paid worker job	Quantitative	Continuous
<code>penpln1</code>	If person is covered by a pension plan connected with job	Qualitative	Nominal
<code>uncoll1</code>	If person was a member of a union/collective agreement	Qualitative	Nominal

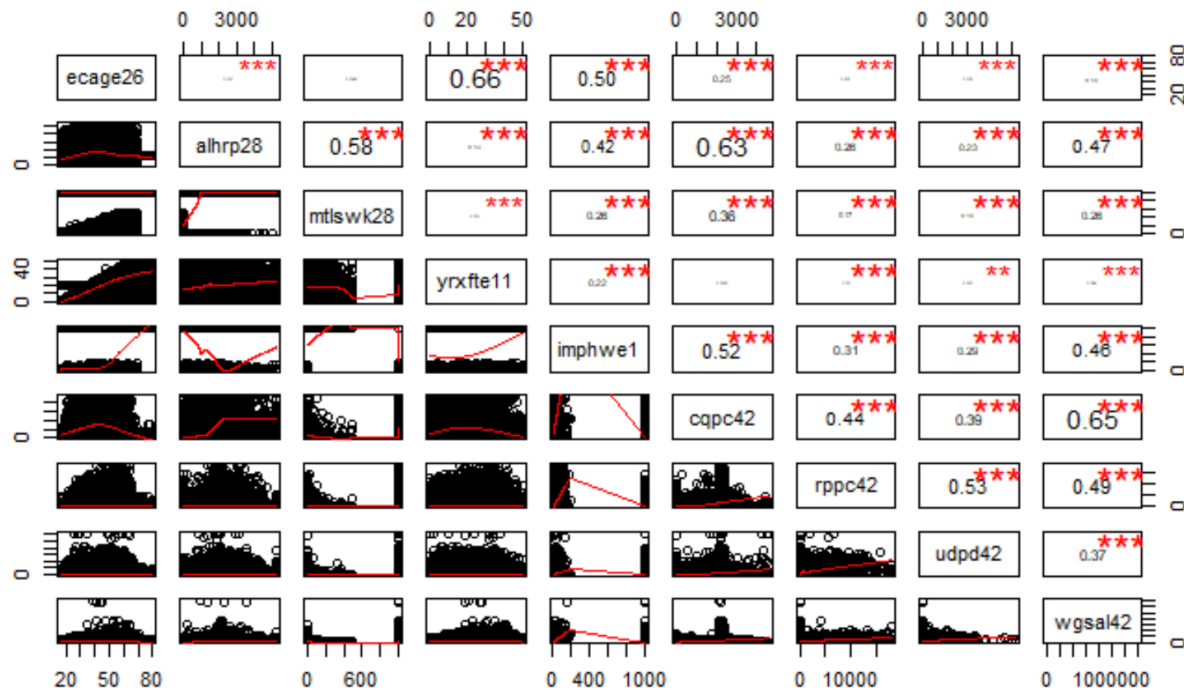
Retirement Income Adequacy and Workplace Pensions
CKME136 Data Analytics: Capstone Course
Melissa Cortina

n07c3g10	Industry code of employer	Qualitative	Nominal
pubpv10	If employer is in the public or private sector	Qualitative	Nominal
cqpc42	CPP/QPP contributions	Quantitative	Continuous
rppc42	Registered pension plan (RPP) contributions/ER-sponsored	Quantitative	Continuous
udpd42	Union dues (both professional premiums)	Quantitative	Continuous
wgsal42	Wages/salaries, before deductions (employment income)	Quantitative	Continuous
hlev18	Highest level of education	Qualitative	Nominal

Of the 26 attributes selected for the analysis, there were 9 quantitative attributes and 17 qualitative attributes. The descriptive statistics for each attribute was calculated and the histograms for each attribute were plotted using WEKA, as per below. It can be noted that two attributes were approximately normally distributed, that is, pvreg25 and pubpv10. Therefore, the final set of selected attributes were balanced/normalized prior to running the algorithms.



Correlation analysis on the quantitative attributes was conducted by calculating the Pearson correlation, as displayed below. All correlation coefficients were below 0.7, therefore, the strength of association between the attributes was not strong. As such, no further selection of attributes was determined from this analysis.



Analysis was also conducted on the qualitative attributes in WEKA by calculating the information gain for each attribute. The attribute that provided zero information gain was mjacg26, and therefore, was removed from the list of selected attributes.

As per the documentation associated with this dataset, for most attributes, an “N/A” entry could be either due to refusal to answer, or a “don’t know” answer, or “not applicable”, and all are represented by numerical codes, usually 97, 98, and 99. Each attribute consisted of its own specific coding, as outlined in the dataset’s documentation. In terms of consistency of the data, since each field/instance contained a numerical value, there were no null values within the dataset. To confirm this, the “apply” function in R was used to sum all N/A entries within each attribute, and all summed to zero.

Further attribute selection was completed depending on if there was enough data within the attribute once the N/A entries were removed. In order to determine this, the documentation provided with the data contained a legend for each attribute. Using this legend, analysis on the “not applicable” entries was conducted, as discussed above. For example, for the attribute reanp1, that is, the reason why a job is not permanent, a “not applicable” entry was coded as “99”. There were 38,951 entries coded as “99” and 4,842 entries coded as “97” which represented the “don’t know” option in the survey. Therefore, a total of 43,793 entries would not contribute to the analysis, approximately 92% of the data. Similar analysis was conducted on the other attributes to determine the final selection of attributes.

Continuing the analysis on the “dummy data” as per above, the “don’t know” answers and “not applicable” answers were replaced with the mode or mean of that particular attribute, depending on whether the attribute was qualitative or quantitative. For example, for the attribute marst26, the marital status of the individual, 71 people out of the total 47,705 answered with “don’t know”. Since this attribute was a qualitative attribute, all these instances were replaced with the mode, which was “1”, that is, the marital status is married. This method was applied to the following attributes as well:

immst15	mtlswk28	penpln1
pvreg25	yrxfte11	uncoll1
dwtenr25	clwkr1	n07c3g10
mortg25	prmjb1	hleveg18
multj28	flprt1	
alhrp28	nocg2e6	

Boxplot analysis in R was conducted on the quantitative attributes, however, instead of removing outliers, if any, for the purposes of this research, the “not applicable” entries were removed.

The results from the correlation analysis, “dummy data” analysis, and analysis of the specific attribute, alfst28, concluded further attribute selection that removed three attributes from the dataset, namely, immst15, mtlswk28, and mjacg26. The final dataset contained 23 attributes and 32,625 instances.

After the data was cleaned using R, the dataset was then converted and imported into WEKA for further analysis. In WEKA, the dataset was transformed from numeric to nominal and balanced.

APPROACHES

The initial approach on this dataset was to analyze a retiree's income and compare it to the salary during their last of employment. If the total retirement income was less than 70% of the income in the last year of employment and the workplace pension was not one of the sources of income, then the industry/occupation that this individual worked for would be targeted by pension plans to offer their employees a workplace pension plan. Unfortunately, the income information for a retiree's last year of income was not clearly outlined, therefore, different approaches were taken in order to analyze the data. The intention of enrolling more members into a pension plan from the perspective of a pension plan was still the focus of this analysis, which, in turn, impacts retirees and their sources of income.

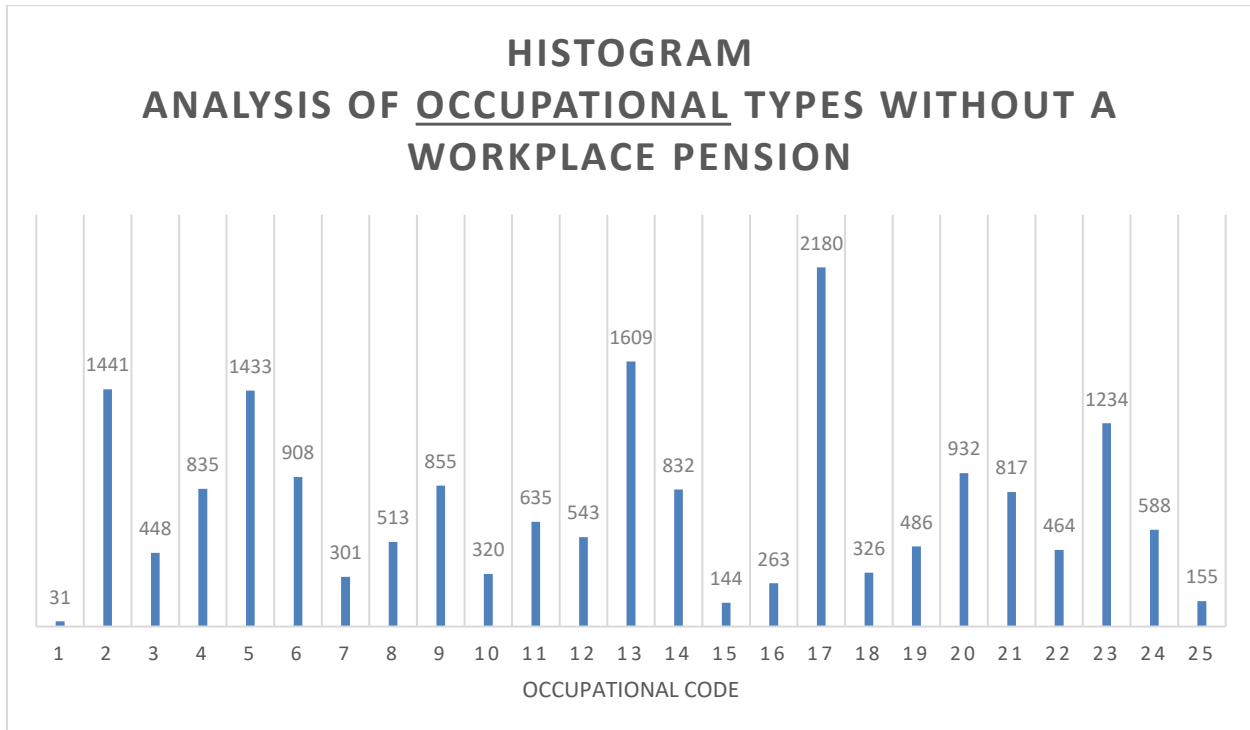
Technique 1: Analysis of Occupational Type and Industry Type

The first approach on the dataset was an analysis of the occupation types and industry types that do not have a workplace pension. A histogram was plotted for each to determine the most common occupation/industry that does not offer a workplace pension.

The dataset was filtered for those instances without a workplace pension plan, that is, when the attribute `penpln1` takes on a value of "2". After filtering the data, two histograms for the attributes for occupational type (`nocg2e6`) and industry type (`n07c3g10`) were plotted. The tables and graphs are illustrated below:

Occupation Code	Description	Count
1	SR Mng	31
2	Oth Mng	1441
3	Prof.	448
4	Admin.	835
5	Cler.	1433
6	Sci.	908
7	Prof. Hlth	301
8	Tech. Hlth	513
9	Soci Sci	855
10	Teach	320
11	ArtSprt	635
12	Buyer	543
13	Retail	1609
14	FoodBev	832
15	ProtServ	144
16	SuppWrk	263
17	Sales	2180
18	Trades	326
19	Cons.	486
20	OthTrade	932
21	Equip	817

22	Trad. Cons	464
23	PrimIndu	1234
24	Machine	588
25	Labourer	155

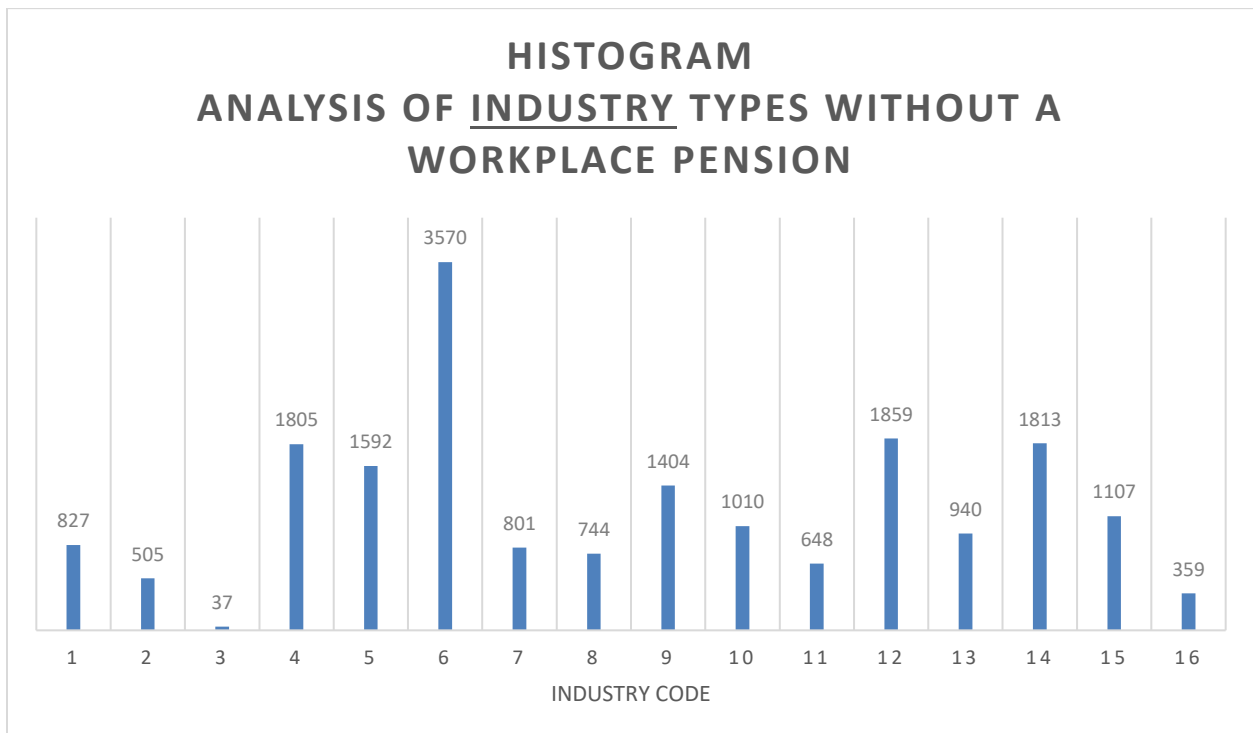


From the histogram plotting the occupation types, it can be observed that the most popular occupation without a workplace pension plan was Sales (code 17). The next popular occupations were Other Types of Non-Senior Management (code 2), Clerical (code 5), and Retail (code 13).

Using this information, a pension plan could target employers who these specific occupation types since majority of those employed in these occupations do not have a workplace pension.

Industry Code	Description	Count
1	Agriculture	827
2	For, fish, min	505
3	Utilities	37
4	Construction	1805
5	Manufacturing	1592
6	Trade	3570
7	Transportation	801
8	Finance, Insur	744
9	Sci and Tech	1404

10	Busi, building	1010
11	Education	648
12	Hlth care, soc assis	1859
13	Infor, cul, rec	940
14	Accom and Food	1813
15	Other Services	1107
16	Pub admin	359



Using the same approach against the industry type, the histogram above illustrates the various industries that do not offer their employees workplace pension plans. It was clear that the Trades industry (code 6), with a count of 3,570, would be an ideal industry to target as a pension plan.

Technique 2: Classification – J48 Decision Tree, Naïve Bayes, K-Nearest Neighbours

The second approach was assigning the attribute `penpln1`, which represented if a person is covered by a pension plan connected with their job or not, as the class attribute, and applying three classification algorithms for supervised learning: J48 Decision Tree, Naïve Bayes, and K-Nearest Neighbours (“KNN”). Instead of targeting one specific occupation or industry, a pension plan could open the option to enroll anyone into the plan, similar to how an insurance company would offer a pension plan. The best model built by one of the classification algorithms would determine if an individual is a good candidate to enroll or not and be used during an application process.

The attribute used as the classifier was `penpln1`, which represented if a person was covered by a pension plan connected with their job or not. An individual who had a workplace pension was represented as a “1”, and therefore, represented a good candidate to be accepted through the application process for the pension plan, and those without a workplace pension were represented by a “2”.

The three classification algorithms used were J48 Decision Tree, Naïve Bayes, and K-Nearest Neighbours. To estimate the skill of the machine learning model, 10-fold cross validation was used for all three algorithms. The details of the results and comparisons between the models can be found in the [Results](#) section.

For the K-Nearest Neighbours model, various values of K were used in order to determine the best possible value of K. As per the confusion matrix results below, specifically for a class value of “1”, the best value of K was 24, which had the best accuracy of 83.84%, while keeping the false positive rate low.

	True Positive Rate (Recall)	False Positive Rate	Precision	ROC Area	Accuracy	Correctly Classified Instances
K = 1	0.789	0.130	0.859	0.879	82.9647%	27,067
K = 5	0.789	0.120	0.868	0.920	83.4489%	27,225
K = 10	0.780	0.107	0.880	0.924	83.6675%	27,297
K = 15	0.779	0.103	0.883	0.925	83.7978%	27,339
K = 20	0.778	0.101	0.885	0.926	83.8185%	27,346
K = 23	0.777	0.101	0.885	0.926	83.8130%	27,344
K = 24	0.777	0.101	0.885	0.926	83.8365%	27,352
K = 25	0.777	0.100	0.886	0.926	83.8166%	27,345

Technique 3: Association Rules – Apriori Algorithm

The last approach was using unsupervised learning on the dataset, specifically the Apriori algorithm, for all those individuals who do have a workplace pension to determine if any further insights could be determined and draw further conclusions on the dataset.

The dataset was filtered for all instances of “1” and the remaining instances were removed. This filtered dataset was then put through the Apriori algorithm using WEKA. The minimum support was set to 0.1, the confidence level was set to 0.9, and the number of rules to be generated was set to 10.

RESULTS

Technique 1

From the histogram for the occupation types, it can be observed that the most popular occupation without a workplace pension plan was Sales (code 17). The next popular occupations were Other Types of Non-Senior Management (code 2), Clerical (code 5), and Retail (code 13).

Using the same approach against the industry type, the histogram illustrates the various industries that do not offer their employees workplace pension plans. It was clear that the Trades industry (code 6), with a count of 3,570, would be an ideal industry to target as a pension plan.

Technique 2

The summary of the confusion matrix calculations for each of the three classification algorithms when the class attribute is equal to “1”:

	J48 Decision Tree	Naïve Bayes	KNN (K = 24)
True Positive Rate (Recall)	0.790	0.846	0.777
False Positive Rate	0.125	0.189	0.101
Precision	0.863	0.817	0.885
ROC Area	0.837	0.920	0.926
Accuracy	83.2516%	82.8310%	83.8365%
Correctly Classified Instances	27,161	27,024	27,352

From the above comparison chart, all three algorithms resulted in similar results. However, it was concluded that the best classification algorithm was K-Nearest Neighbours when K = 24 for this dataset. KNN had the best accuracy of 83.84%, while keeping the precision high and false positive rate low. The precision was the best for KNN versus the other two models, and therefore, KNN had the lowest recall rate. To further support the choice of the KNN model, the area under the receiver operating characteristic curve (“ROC Area”) was the greatest for the KNN model, which tells us the probability of classifying an instance correctly was high.

Technique 3

The last technique was applying association rules, specifically, the Apriori algorithm. After seven cycles being performed to obtain ten rules, nine of the ten rules contained the attribute flprt1 (if a person’s job was fulltime or not) with a value of “1”, that is, the person held a fulltime position. Six of the ten rules contained the attribute multj28 (if a person held multiple jobs throughout the year or not) with a value of “2”, that is, the person did not hold multiple jobs. Therefore, it can be concluded that the most common characteristics of an individual who would be a good candidate for a pension plan to enroll is a fulltime worker who holds one job throughout the year.

CONCLUSION

A retiree's income should be approximately 70% of their income in their last year of employment and one of the components of this income is the pension's accrued pension from their workplace during employment. Therefore, pension plans should seek out those employers who not offer their employees pension plans, not only for the aid of when an employee retires but also to sustain the pension plan through continuous new enrollments.

Using the Canadian data from the Survey of Labour and Income Dynamics for 2011 from Statistics Canada, various techniques were applied to gain insight into the working population and their workplace pensions. Three techniques were applied: analysis of occupation type and industry type for all those who did not have a workplace pension, supervised learning using three classification algorithms, and unsupervised learning by applying the Apriori algorithm.

From the analysis, a pension plan could target employers of a specific occupation or industry to offer their employees a workplace pension. The occupation to be targeted based on the dataset was a position in Sales and the industry to be targeted was the Trades industry.

If a pension plan opened their plan to any individual, the application process could use the model built using K-Nearest Neighbours with a K value of 24, since it was the best-suited algorithm for supervised learning for this dataset.

Further insights were captured when the Apriori algorithm was applied for unsupervised learning. It was concluded that those who have a fulltime job and do not hold multiple jobs throughout the year are the best-suited candidates to enroll into a pension plan.

To continue the analysis on this dataset for further investigations and insights, other classification models, such as, logistic regression, could be also be applied and compared to the existing models studied. In addition, to further analyze the current models, other techniques for cross-validation to analyze the models could be applied, such as, splitting the data into training and testing sets. Lastly, analysis on the attributes there were excluded during the first selection of attributes could be studied to determine a different combination of characteristics.

REFERENCES

- Baldwin, Bob. "Research Study on the Canadian Retirement Income System." *Ministry of Finance, Government of Ontario*, Nov. 2009, www.fin.gov.on.ca/en/consultations/pension/dec09report.pdf.
- Baldwin, Bob. "Assessing the Retirement Income Prospects of Canada's Future Elderly: A Review of Five Studies." *C.D. Howe Institute*, ISBN 978-0-88806-980-1, Sep. 2016, https://www.cdhowe.org/sites/default/files/attachments/research_papers/mixed/Commentary%20456_0.pdf.
- Baldwin, Bob. "The Evolving Wealth of Canadians Approaching Retirement." *C.D. Howe Institute*, ISBN: 978-1-989483-00-8, 28 Mar. 2019, https://www.cdhowe.org/sites/default/files/attachments/research_papers/mixed/Working%20Paper%200328_web.pdf.
- Baldwin, Bob, and Richard Shillington. "Unfinished Business: Pension Reform in Canada." *Institute for Research on Public Policy*, no. 64, June 2017, <https://irpp.org/wp-content/uploads/2017/06/study-no64.pdf>.
- Cross, Philip. "The Reality of Retirement Income in Canada." *Fraser Institute*, April 2014, <https://www.fraserinstitute.org/sites/default/files/reality-of-retirement-income-in-canada.pdf>.
- MacDonald, Bonnie-Jeanne. "Filling the Cracks in Pension Coverage: Introducing Workplace Tax-Free Pension Plans." Toronto, ON: *National Institute on Ageing White Paper*, 2019, <https://www.ryerson.ca/nia/white-papers/filling-the-cracks-in-pension-coverage.pdf>.
- MacDonald, Bonnie-Jeanne, Osberg, Lars, and Moore, Kevin D. "How Accurately Does 70% Final Employment Earnings Replacement Measure Retirement Income (In)Adequacy? Introducing the Living Standards Replacement Rate (LSRR)." *Cambridge University Press, Astin Bulletin* 46(3), 627-676. doi: 10.1017/asb.2016.20, 2016.
- MacQueen, Alexandra. "Not so fast, not so easy: Retiring on a low income is a challenge in Canada." *The Globe and Mail*, 19 Jan. 2019, www.theglobeandmail.com/investing/personal-finance/retirement/article-not-so-fast-not-so-easy-retiring-on-a-low-income-is-a-challenge-in/.
- McCarthy, Shawn. "Many Canadians entering retirement with inadequate savings, study says." *The Globe and Mail*, 16 Feb. 2016, <https://www.theglobeandmail.com/globe-investor/retirement/retire-planning/many-canadians-entering-retirement-with-inadequate-savings-study-says/article28761394/>.
- Statistics Canada. Survey of Labour and Income Dynamics, 2011 [Canada]: Person File [Public-use Microdata File]. Ottawa, Ontario: Statistics Canada. Income Statistics Division [Producer and Distributor], 2011.

Vettese, Fred. "Why Canada Has No Retirement Crisis." *Rotman International Journal of Pension Management*, Volume 6, Issue 1, 2013,

<https://www.morneaushepell.com/sites/default/files/documents/336-why-canada-has-no-retirement-crisis-fred-vettese/rotman-whycanadahasnoretirementcrisis.pdf>.