Performance analysis of Ontario Works in Toronto

ABSTRACT

In this project, we propose the methodology and models for analysing the performance of Ontario Works in Toronto over the last decade. The public perception of Ontario Works, a provincial social assistance program is that it is not a very efficient process to help people get back on their feet after accidents or unemployment. The consensus seems to be that people are using Ontario Works as a long-term support system rather than for it's intended purpose. Our solution is to find an appropriate metric to identify groups of people from the household level census data who have stayed in the assistance program without ever exiting welfare and build models/classifiers to identify such groups. Using the census data from Toronto's Open Data Catalogue, we identify the attributes of groups staying in the assistance program.

1.INTRODUCTION

Ontario Works (OW) benefits are for people who need money because they cannot find work or temporarily cannot work. It has two parts: Financial assistance and Employment assistance. The intent of the Ontario Works program is to help people in temporary financial need find sustainable employment and achieve self-reliance through the provision of effective, integrated employment services and financial assistance [1]. The program serves about 450,000 beneficiaries in the province on an average month [2]. This program costs $10 billion annually.

Reform of the Ontario Works has always been on the agenda of the campaign of political parties [3]. The focus of the conversation has been on streamlining the program and detecting fraud. The steps to prevent fraud include: establishing a Welfare Fraud Hotline to detect and prevent fraud; introducing new legislation to expand the investigative powers for eligibility review staff; setting up a province-wide fraud control database to monitor and track results of welfare fraud investigations; and implementing information-sharing agreements with both provincial and federal governments [4].

Toronto city, the capital of Ontario is home to one-fourth of Ontario's population. As a part of the open data initiative of the city, the public has access to all the anonymized survey data. With the availability of the "Employment & Social Services" open data, the performance of Ontario Works can be studied in detail. Toronto city government has provided the census data with the information of the assistance program at the Head of Household level dataset from the time period 2004 to 2018. It provides household information with reference to the program [5]. The count of total cases, new cases and closed cases per month for different groups of people across Toronto is available in the data.

In this project, we focus on identifying the groups which have stayed in assistance for a long time. This is measured by the number of cases closed (Exit cases) over the decade for a certain group. The data has all the required demographic information to classify the population of Toronto. The dependence of a population group on assistance, the number of new people joining the program and the number of people exiting the program are all useful metrics for the analysis. A group of people are said to be been in the system for a long time if they have zero Exit cases over the given time period. 'EXIT BINARY' is a binary column with value as 1 when there have been Exit cases and 0, otherwise. This is the outcome variable in the analysis.

We propose different algorithms and features to identify the groups of people who have never exited the system (Exit cases =0). The classification models used are decision tree and logistic regression. These models when provided with the different attributes of a group such as age, education and gender will classify them into different categories. In our analysis, the categories are: Exit cases is 0, stayed in the system without ever exiting or Exit cases is not 0, exited the social assistance program at least once. The accuracy of the classification algorithms is the fraction of correct predictions out of the total predictions. The higher the accuracy of a model, the better it is at classifying the groups of people. The confusion matrix is used to further investigate the success of a classification model.

Alternatively, Apriori association rule mining algorithm could be used for performing the same analysis. It consists of two procedures: First, finding the frequent itemset in the database using a minimum support and constructing the association rule from the frequent itemset with specified confidence [6]. Support is the proportion of transaction in which an itemset appears. Confidence is the proportion of transactions with item X, in which item Y also appears. For a given class value of the explained outcome variable, EXIT BINARY, this model can build a set of rules satisfying the pre-established support and confidence levels. For example, for EXIT BINARY = 0 i.e., the group of population that has never exited Ontario Works, the rule is Gender = "Female" and Education level = "High School" , given that the rule has met the provided support and confidence levels.

The different classifiers and association rule mining algorithms are applied on the census data to identify the groups of population with EXIT Binary as 0 and 1. The remainder of this project focuses on understanding the census data, EDA, identifying the features, building the listed models, measuring their accuracy and validation scores and interpreting the results.

As mentioned earlier, Ontario spends $10 billion annually on the social assistance program. Various measures are in place to detect fraud. The incentive to detect fraud is very high, as benefits will continue to reach undeserving people every month till the fraud is detected. There are future savings in detecting and stopping welfare fraud. However, Ontario Works neither has the budget or personnel to focus as much as required on identifying the fraudulent cases. A model or system that helps the workers identify the groups of population prone to staying in the system longer.i.e. Groups with EXIT BINARY =0 will focus the fraud detection efforts. Instead of scouring the entire population for fraudulent cases, workers can look at the results of the model and increase their efforts on such groups. Currently, the program is one year behind on investigations of approximately 6000 fraud tips. The model will help in deciding which among those cases (based on the person's attribute) has a higher chance of being a fraud.

2.RELATED WORK

Ontario Works (OW) prepares regular monthly and yearly reports like any other organisation to measure their performance and KPIs [7]. These reports are mostly diagnostics in nature and focus on the volume of the workload and demographics of the beneficiaries. Internal reports or analysis of OW may exist , but are not readily available to access.

As OW comes under the governance of Ministry of Social Services, the managing body is responsible for publishing their yearly goals, targets, concerns, current performance and recommendations to the public domain [8]. From this report, we understand that the ministry has acknowledged existence of rampant welfare fraud and is in lookout for recommendations to tackle this issue. Our project is a step in the right direction for doing this.

However, independent research by academia does exist on this subject. The most similar research to our topic being the PhD thesis of Vaillancourt [9]. It explores how OW was redeveloped in 1997 for Northern Ontario keeping in mind the then prevailing social and economic conditions. The analysis was focussed on the effectiveness of welfare program in regions with low Human Capital Index (less employable skills). The commonality of the two analysis is that the focus is on Ontario works but the nature and the aim of the two studies are vastly different. Our analysis solely focuses on detecting the groups of people who have not exited welfare system for a long time.

The data for the analysis was procured from the open data catalogue of Toronto city. Government agencies and cities have lately realised the value of providing open data for researchers. As part of one such initiative Ontario Social Assistance Database (OSAD) was made available to the public [10]. The data contained family, member, pay detail, income and skill information of people in the OW program. 10 publications had been made using this data [11]. Most studies focussed on a certain demographic and its specific problems or trends. Our analysis varies focuses on identifying and classifying the different sects of population based on their past exits from OW.

The study closest to our analysis in terms of aim of the study is the study focussing on the leavers of OW by Lightman [12]. The study data was collected solely from leavers of OW and focussed on identified the patterns in those subjects. Our analysis also focuses on the same objective, but the data used, and approach used is completely different. For our analysis, census data of the general population with a multitude of features is used but in the other research, data was collected on telephonic surveys and did not capture a lot of the subject's attributes. The new patterns and insights developed are shared in the remainder of the paper.

3.DATA

The data was taken from the Ontario Social Assistance Database (OSAD). The data is made available to the public as a part of City of Toronto's Open Data Catalogue. The data available is at an aggregated format and has no personal information that can be traced back to an individual. Each entry on this data pertains to a subset of the population availing OW benefits.
The different columns in the data provide information about the attributes of the population such as their age, education, immigration status, gender etc. There is no unique identifier in the data.

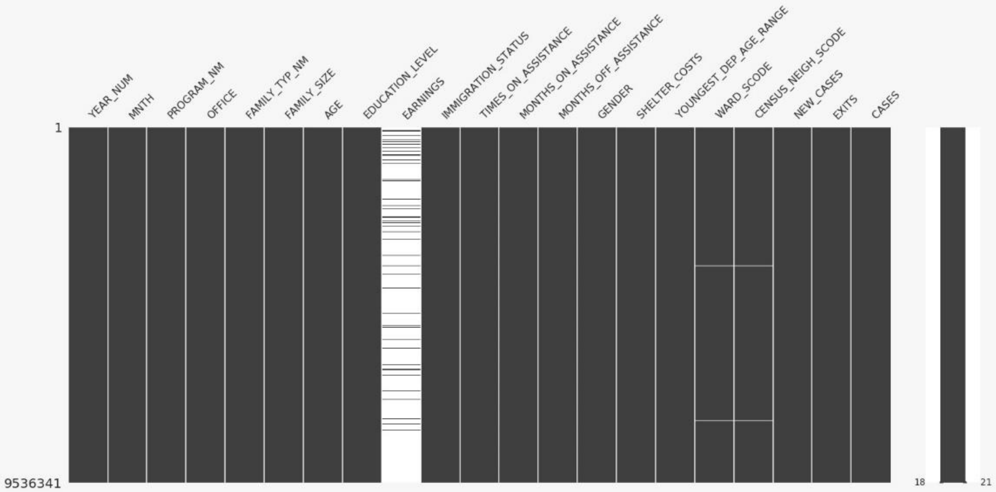
The data is available at month level from the time period of January 2004 to October 2014.
The dataset has 9.5 million rows and 21 columns. Out of the 21 columns, 18 columns provide information about the population using OW. The data has 3 numerical columns: NEW_CASES, EXITS and CASES. These are the count of cases for the selected population in a given month.

Table 1 : Data Dictionary

Column Name	Column Description
YEAR_NUM	Year
MNTH	Month
PROGRAM_NM	Toronto Employment and Social Services Office name
OFFICE	Name of Program
FAMILY_TYP_NM	Type of family structure
FAMILY_SIZE	Number of members on Ontario Works (OW) in a family unit
AGE	Age range of OW head of household
EDUCATION_LEVEL	Highest level of education
EARNINGS	The amount earned by Head of Household(HOH) for the period
IMMIGRATION_STATUS	The immigration status of the HOH

TIMES_ON_ASSISTANCE	No of times a head of household has been granted OW since 2002. Zero means the resident has not been on OW in Toronto before, 4+ means more than 4 times.
MONTHS_ON_ASSISTANCE	The number of months a resident has been on OW from Grant date to exit date or period start date (Current date)
MONTHS_OFF_ASSISTANCE	The number of months the resident has not been receiving OW
GENDER	Male or Female
SHELTER_COSTS	Cost ranges of shelter expenditure in a household
YOUNGEST_DEP_AGE_RANGE	Age ranges of the youngest child in a household
WARD_SCODE	LIT defined wards of Toronto
CENSUS_NEIGH_SCODE	LIT defined Toronto neighbourhoods
NEW_CASES	The number of new cases
EXITS	The number of cases that are not receiving OW anymore
CASES	The number of total type of cases

For the purpose of data cleaning, each dimension of the dataset was investigated individually for missing, null or junk values. EARNINGS column had more than 90% of its values as blank, so that column could not be part of the future analysis. Since, the data collected was census data and the information missing were categorical variables, there were no viable ways to treat such missing. Such columns were not considered for further analysis.



EDUCATION_LEVEL columns has records with “UNKNOWN” as the category. Data with such values cannot be used for building classifiers or models as it makes interpretation of results cumbersome. Data cleaning was performed in case of FAMILY_SIZE. The possible values for this column are 1,2,3,4 and 4+. However, there were two different formats of the value ‘1’ present in this column. So, data in this column was sanitized. Data transformation was not required as most of the columns were categorical. PROGRAM_NM has only one value – ‘Ontario Works’.

In this analysis, the focus is on aggregated EXITS for a population group over the given time period. So, data at a YEAR_NUM – MNTH level is not required, instead the data is aggregated.
WARD_SCODE and CENSUS_NEIGH_SCODE provide the geography information. This analysis is focussed on the population demographics of Toronto overall. This level of information is not required, and these columns are not considered for our analysis.

Currently, the data is available at a very granular level. For the purpose of the analysis, after removing the unwanted columns, data is aggregated at the remaining levels. The aim of the analysis is to identify the groups of people who have never exited the OW program.i.e. EXITS = 0. The focus is not on predicting the number of EXITS (people leaving OW) but rather on identifying if a group has ever exited OW or not. The outcome variable is a derived metric of EXITS called EXIT_BINARY. It is 1 when a group has EXITS > 0 and 0 when a group has EXITS = 0.

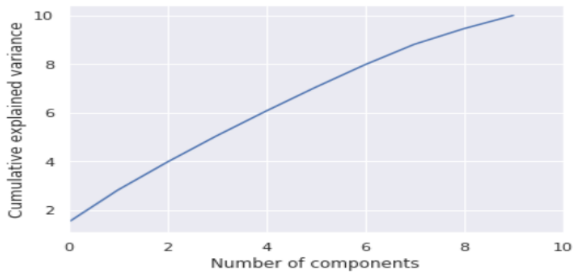
There is a need for aggregation of data after dropping a column because when one attribute of a population is removed, the size of the population increases. For example, data is at AGE-GENDER level and EXITS is the sum of cases for this level of data. Now, if the AGE column in dropped, the population is at GENDER level and the EXITS must be aggregated to reflect the number of exit cases at GENDER level. In this new aggregated data, based on the EXITS, calculate the outcome variable, EXIT_BINARY.

After the data clean up process, the data is aggregated at this level OFFICE - FAMILY_SIZE – AGE - IMMIGRATION_STATUS - TIMES_ON_ASSISTANCE - MONTHS_ON_ASSISTANCE - MONTHS_OFF_ASSISTANCE - SHELTER_COSTS - GENDER - YOUNGEST_DEP_AGE_RANGE.
This is the data for our further analysis. Get the sum EXITS at this level of data and calculate the outcome variable EXIT_BINARY.

4.RESULTS
4.1.FEATURE SELECTION

We have 21 features to predict our outcome. In order to do classification and association mining, we need to have best features that can explain our outcome variable. For reducing the dimensionality of our large dataset, we performed Principal Component Analysis (PCA) to transform our large set of variables into a smaller one while still preserving most of the information. After data cleaning and transformation, we had 10 predictors that were found suitable to go ahead with our analysis. PCA results showed that 85% variance is explained by 8 out of the 10 features. We identified these 8 features from PCA and used them as predictors in classification and association mining.

Cumulative explained variance of the features



4.2. CLASSIFICATION

4.2.1. DECISION TREE

The purpose of a decision tree algorithm is to classify the population if they have exited or never exited the Ontario social assistance program. We are classifying using an outcome variable - EXIT_BINARY where 0 means that population has never exited the assistance program and 1 means that population has exited the program.

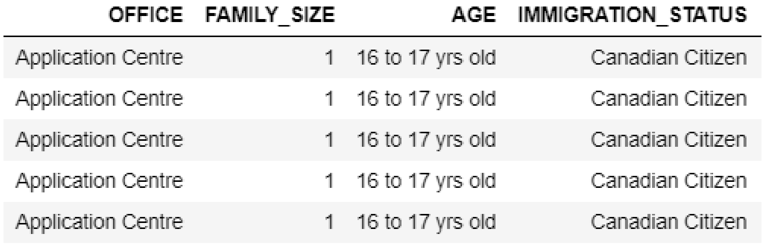
Predictor variables - FAMILY_SIZE, AGE, TIMES_ON_ASSISTANCE, MONTHS_OFF_ASSISTANCE, SHELTER_COSTS, OFFICE, IMMIGRATION_STATUS, GENDER

Predictors with one-hot encoding - FAMILY_SIZE_1, FAMILY_SIZE_2, FAMILY_SIZE_3 ………..
…GENDER_MALE, GENDER_FEMALE.

Outcome variable- EXIT_BINARY, 0 when population has never exited OW and 1 otherwise.

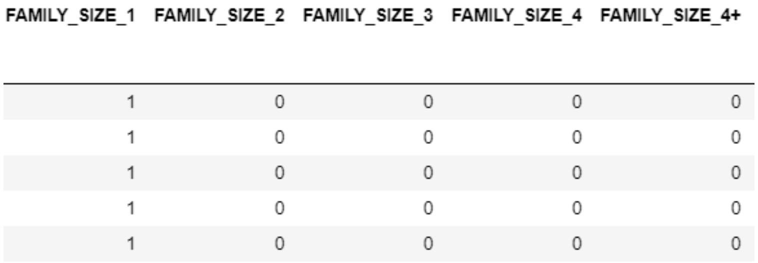
Our predictor variables are categorical. Hence, one-hot encoding was done to convert categorical variables to binary form. If this was not done and label encoding done instead, the data would be treated as numerical variable. For example, GENDER is a categorical variable with two values Male and Female. When one-hot encoding is done, two binary columns are created: GENDER_Male and GENDER_Female. For records with GENDER=" Male", the first column has values =1 and 0, otherwise.

Table1. Predictor variable without one-hot encoding



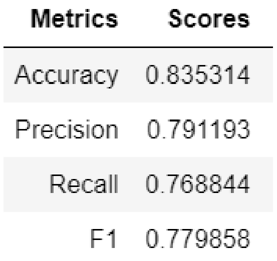
OFFICE	FAMILY_SIZE	AGE	IMMIGRATION_STATUS
Application Centre	1	16 to 17 yrs old	Canadian Citizen
Application Centre	1	16 to 17 yrs old	Canadian Citizen
Application Centre	1	16 to 17 yrs old	Canadian Citizen
Application Centre	1	16 to 17 yrs old	Canadian Citizen
Application Centre	1	16 to 17 yrs old	Canadian Citizen

Table2. Predictor variable with one-hot encoding



FAMILY_SIZE_1	FAMILY_SIZE_2	FAMILY_SIZE_3	FAMILY_SIZE_4	FAMILY_SIZE_4+
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0

Now, Data is divided into train and test set. The parameter test size was given value 0.2; meaning test sets will be 20% of the dataset & training dataset's size will be 80% of the entire dataset. Following that decision tree was fitted on training data, predicting labels for validation dataset and providing different performance metrics of the model. Cross validation was applied to determine the number of trees that give the better accuracy score. Using hit and trial approach, at the maximum depth =10 the decision tree provided us best accuracy score 83.5. We further calculated confusion matrix, precision, recall and f1 score for the model. Here are the results of classification report:



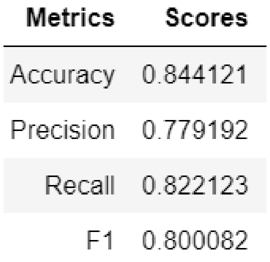
Metrics	Scores
Accuracy	0.835314
Precision	0.791193
Recall	0.768844
F1	0.779858

83% of the predictions were made correctly. 79% of the all the positive predictions made turned to be correct. Recall is 77%, the ratio of correct positive predictions to total positive observations. In some cases, there is a trade off between precision and recall. So, the better metric to use in F1 score, the harmonic means of both. Harmonic mean is used to punish extreme values of one of the metrics. F1 score can be maximised with the optimal balance of precision and recall.

4.2.2 DECISION TREE WITH RANDOM FOREST

After building a model using decision tree classifier, we decided to build a model that can give better f1 score for the given decision tree. Through a better f1 score we would seek a balance between Precision and Recall and uneven class distribution (large number of actual negatives). The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. We have performed hyper-parametric tuning using Grid search CV to achieve our best model. Our model was able to achieve 84.4% accuracy and 80% f1 score. The parameters that were given are - n_estimators = 100, criterion = 'gini', max_depth = 15, bootstrap=False, random_state=43.

Results of random forest classification report –



Metrics	Scores
Accuracy	0.844121
Precision	0.779192
Recall	0.822123
F1	0.800082

Accuracy of predictions is 84%. Of all the positive predictions made 78% turned to be correct. 82% of the actual positives were identified correctly. F1 score = 80% means that the model was getting good balance of precision and recall was reached.

4.2.3. LOGISTIC REGRESSION

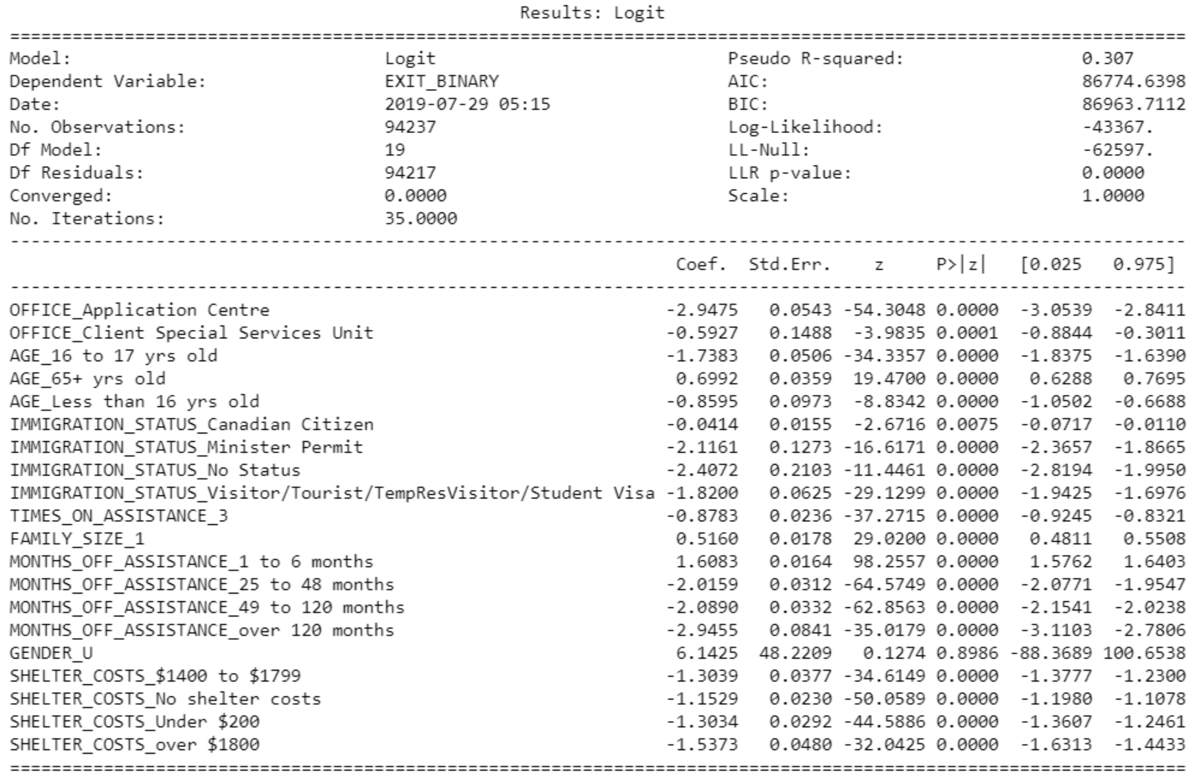
The purpose of the regression model is to predict using the given predictors the class of the outcome variables (population will exit the program or not). Our outcome variable is binary where 0 means that population has never exited the program and 1 means that population has exited the program.

Predictor variables - FAMILY_SIZE, AGE, TIMES_ON_ASSISTANCE, MONTHS_OFF_ASSISTANCE, SHELTER_COSTS, OFFICE, IMMIGRATION_STATUS, GENDER

Predictors using dummy encoding - AGE_16 to 17 yrs old, AGE_65+ yrs old, AGE Less than 16 yrs old ……SHELTER_COSTS_Under 200','TIMES_ON_ASSISTANCE_3'

Outcome variable- EXIT_BINARY, 0 when population has never exited OW and 1 otherwise.

To build our logistic regression model we used the same 8 features as our predictor variables that we had identified using PCA. Firstly, we converted the features to dummy variables as they were categorical. Now, the number of predictor variables were converted to 56 variables. We further used recursive feature elimination (RFE) technique to identify any weak feature out of these 56 variables. 35 out of 56 variables were eliminated using RFE. We them implemented our model using the remaining 21 variables. Following is the result of our implemented model:



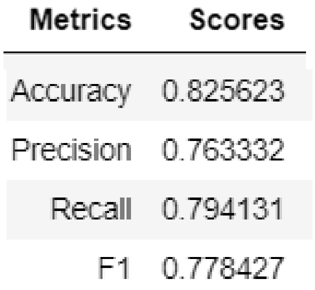
```
                                                 Results: Logit
=================================================================================================================
Model:                              Logit                            Pseudo R-squared:                 0.307
Dependent Variable:                 EXIT_BINARY                      AIC:                              86774.6398
Date:                               2019-07-29 05:15                 BIC:                              86963.7112
No. Observations:                   94237                            Log-Likelihood:                   -43367.
Df Model:                           19                               LL-Null:                          -62597.
Df Residuals:                       94217                            LLR p-value:                      0.0000
Converged:                          0.0000                           Scale:                            1.0000
No. Iterations:                     35.0000
-----------------------------------------------------------------------------------------------------------------
                                                                Coef.  Std.Err.    z     P>|z|   [0.025   0.975]
-----------------------------------------------------------------------------------------------------------------
OFFICE_Application Centre                                     -2.9475   0.0543 -54.3048 0.0000  -3.0539  -2.8411
OFFICE_Client Special Services Unit                           -0.5927   0.1488  -3.9835 0.0001  -0.8844  -0.3011
AGE_16 to 17 yrs old                                          -1.7383   0.0506 -34.3357 0.0000  -1.8375  -1.6390
AGE_65+ yrs old                                                0.6992   0.0359  19.4700 0.0000   0.6288   0.7695
AGE_Less than 16 yrs old                                      -0.8595   0.0973  -8.8342 0.0000  -1.0502  -0.6688
IMMIGRATION_STATUS_Canadian Citizen                           -0.0414   0.0155  -2.6716 0.0075  -0.0717  -0.0110
IMMIGRATION_STATUS_Minister Permit                            -2.1161   0.1273 -16.6171 0.0000  -2.3657  -1.8665
IMMIGRATION_STATUS_No Status                                  -2.4072   0.2103 -11.4461 0.0000  -2.8194  -1.9950
IMMIGRATION_STATUS_Visitor/Tourist/TempResVisitor/Student Visa -1.8200   0.0625 -29.1299 0.0000  -1.9425  -1.6976
TIMES_ON_ASSISTANCE_3                                         -0.8783   0.0236 -37.2715 0.0000  -0.9245  -0.8321
FAMILY_SIZE_1                                                  0.5160   0.0178  29.0200 0.0000   0.4811   0.5508
MONTHS_OFF_ASSISTANCE_1 to 6 months                            1.6083   0.0164  98.2557 0.0000   1.5762   1.6403
MONTHS_OFF_ASSISTANCE_25 to 48 months                         -2.0159   0.0312 -64.5749 0.0000  -2.0771  -1.9547
MONTHS_OFF_ASSISTANCE_49 to 120 months                        -2.0890   0.0332 -62.8563 0.0000  -2.1541  -2.0238
MONTHS_OFF_ASSISTANCE_over 120 months                         -2.9455   0.0841 -35.0179 0.0000  -3.1103  -2.7806
GENDER_U                                                       6.1425  48.2209   0.1274 0.8986 -88.3689 100.6538
SHELTER_COSTS_$1400 to $1799                                  -1.3039   0.0377 -34.6149 0.0000  -1.3777  -1.2300
SHELTER_COSTS_No shelter costs                                -1.1529   0.0230 -50.0589 0.0000  -1.1980  -1.1078
SHELTER_COSTS_Under $200                                      -1.3034   0.0292 -44.5886 0.0000  -1.3607  -1.2461
SHELTER_COSTS_over $1800                                      -1.5373   0.0480 -32.0425 0.0000  -1.6313  -1.4433
=================================================================================================================
```

Pseudo R-squared (McFadden's R squared) is at a respectful level of 36%. Of the listed predictors, except GENDER_U (“Unknown”) are statistically significant as their $p < 0.05$. MONTHS_OFF_ASSISTANCE = “over

120 months” and OFFICE = “Application centre” have the highest absolute values of the predictor coefficients. Hence, have the maximum impact on the outcome variable.

Next, we did validation of the model. Data slicing was done to divide data into training and testing dataset. Parametric size=0.3 was set. Logistic model was fitted on the training dataset, predicting labels for validation dataset. Using 10-fold cross validation, we achieved accuracy of 82%. We then calculated confusion matrix, precision, f1 score and recall.

Results of the classification report



Metrics	Scores
Accuracy	0.825623
Precision	0.763332
Recall	0.794131
F1	0.778427

Lower values of all the performance metrics suggest that logistic regression does not suit this dataset as well as the other methods.

4.3. ASSOCIATION RULE MINING

Association rule mining is one of the ways to find patterns in data. It finds the features that occur together for a value of class variable. In our analysis, the outcome variable, EXIT_BINARY and class values are 0 and 1. The focus is on finding the features that occur together when EXIT_BINARY = 0. This value is 0 for the groups of population which have never exited the OW program.

Features selected by PCA are used for association mining. In the dataset, filter the records with EXIT_BINARY =0. This is the value of class variable we are focussed in this analysis. For the population of people staying in OW program (EXIT_BINARY =0), identify their frequently occurring features. By knowing these features, the fraud detection efforts can be focused on this particular group of people as they don’t exit the OW system at all. These are the groups of people who should be checked first when complaints are registered with OW fraud.

There are 8 selected features (all categorical variables) with each having varying number of categorical variables. One-hot encoding must be done on all the categorical columns before using the Apriori association mining algorithm. For example, GENDER is a categorical variable with two values Male and Female. When one-hot encoding is done, two binary columns are created: GENDER_Male and GENDER_Female. For records with GENDER=”Male”, the first column have values =1 and 0,otherwise.

The strength of an association can be measured with support and confidence. Support is a measure of the popularity of an itemset. For example, if 5 of the 10 groups have Gender = “Female”, the support is 0.5. The more frequent an itemset occurs, the higher the support. Confidence is the proportion of transactions with an item X, in which item Y also appears.

In this scenario, multiple iterations of Apriori algorithm is run with different values for support and confidence to find the frequent behaviour in population of people who don't exit OW.
The Apriori algorithm is run on the filtered dataset: EXIT_BINARY =0, people who have been in OW without ever exiting. All records in this dataset have EXIT_BINARY = 0.

Support (EXIT_BINARY is 0, No exits) = 1
Confidence (EXIT_BINARY is 0 -> Any feature of population) =
Support (feature of population, EXIT_BINARY is 0)/1 = Support (feature of population)

This is a special case when support is always 1 and hence, confidence(feature) = support(feature). This happens as we have filtered the dataset with the required class variable separately. Support and confidence condition will have the same value as explained above.

Support = 0.25 and Confidence = 0.25

Running the Apriori algorithm with the mentioned support and confidence gives the below result. The frequent item sets are listed below.

support	**Item set**
0.26	AGE_40 to 49 yrs old
0.26	AGE_30 to 39 yrs old
0.28	MONTHS_OFF_ASSISTANCE_7 to 24 months
0.31	TIMES_ON_ASSISTANCE_1
0.31	IMMIGRATION_STATUS_Permanent Resident
0.31	GENDER_F', 'IMMIGRATION_STATUS_Canadian Citizen
0.32	TIMES_ON_ASSISTANCE_4+
0.35	FAMILY_SIZE_1
0.44	GENDER_M
0.55	IMMIGRATION_STATUS_Canadian Citizen
0.56	GENDER_F

Different models with varying values of support an confidence cannot be built in this case.

In the case of most frequent item set, Support (GENDER is F) = 0.56
Support (EXIT_BINARY is 0 or No exits in population) = 1

Confidence (Gender = F | EXIT_BINARY is 0) = Support (GENDER, EXIT_BINARY is 0)/Support (EXIT_BINARY is 0) = 0.56/1 = 0.56

Therefore, the support and confidence are always same in this case. The purpose of experiment is to identify the features of the population that have stayed in OW with no exits. The sets of population satisfying the above listed rules have a higher chance of staying in OW without exits.
Association mining is not predictive, it's inferential. It assumes that trends of past will continue. So, validation is not used for association mining as it is used in prediction models. The confidence and support levels mentioned are the validation required.

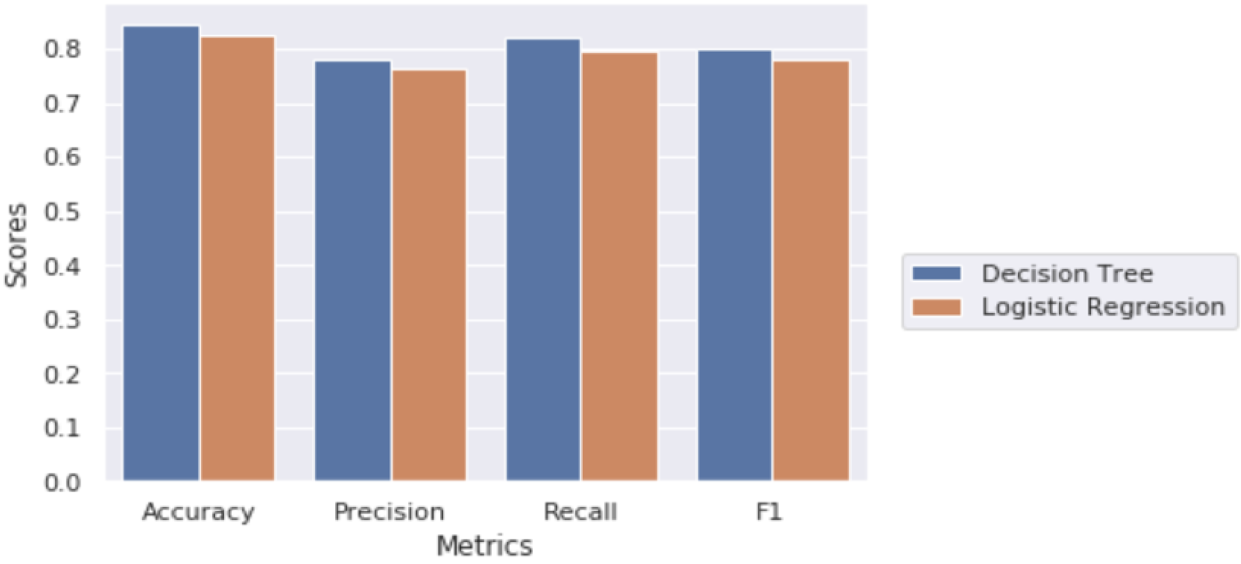
5.CONCLUSION

The focus of the analysis was on identifying the different features associated with the population that does not exit OW. There were 18 different attributes of the population in the census data. With the features finalised using Principal Component Analysis (PCA), classifiers were built to show the characteristics of the population that play an important role in deciding behaviour of population.

Firstly, the classifier built was decision tree. We concluded that this classifier algorithm performed well in terms of all metrics like accuracy, precision, recall and f1 score. But the model still had scope for improvement. So, random forest algorithm was used to reduce prediction error.

Secondly, the classifier built was decision tree with random forest algorithm incorporated. This was the best classifier model out of the lot. It was an expected result as a large number of trees (15) operating as a committee (outcome is decided by the majority vote) will always outperform any of the individual models.

Lastly, the classifier was logarithmic regression. The results were very close to those of the previous model. As mentioned above, the regression coefficient shows us both the order and magnitude of the importance of each predictor in the final result.



In association rule mining, Apriori algorithm was used. The dataset was filtered for EXIT_BINARY=0 .i.e. population without ever exiting OW. This structure ensured that the support for EXIT_BINARY always remained as 1. Confidence for a predictor given that EXIT_BINARY =0 is same as the support for the predictor. This special case of data helped me understand the nuance of Apriori algorithm. GENDER = Female had the highest support, but this could be attributed to the fact that the gender column had only two categories to start with. So, the occurrence of the different genders was high in the data compared to other categorical variables with more categories.

Ontario Works has realised the importance of fraud detection. With their limited personnel and resources, they are a year behind the complaints regarding welfare fraud. Helping them understand that certain population groups have higher chance of defrauding OW based on their past behaviour is the first step. Once this point is put across, the results of the classifiers and models can be discussed in detail with Ontario Works. This direction provided by us from the analysis will help them focus their efforts and in turn save welfare money in long term. Understanding the demographics of groups that stay in welfare longer will help in policy revision and changes to the current OW setup. OW must put in additional focus on these vulnerable population fringes to reduce their dependence on Welfare.

6.REFERENCES

Ministry of Children, Community and Social Services
1.https://www.mcss.gov.on.ca/en/mcss/programs/social/ow/

2.https://www.mcss.gov.on.ca/en/mcss/open/sa/owCaseLoadReport.aspx

CBC news –

3.https://www.cbc.ca/news/canada/toronto/doug-ford-welfare-social-assistance-ontario-works-odsp-1.4885584

Ministry of Community, Family and Children's Services

4.http://www.ontla.on.ca/library/repository/mon/6000/10313026.pdf

Toronto Open Data Catalogue -
5.https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#15580d71-31ca-010e-3895-6b77b49d966f

Apriori Algorithm – "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation" by Sheila A. Abaya, ISSN 2229-5518 –
6.https://pdfs.semanticscholar.org/4dcf/50517e96401800a32cdd1c39392d8e4ccda8.pdf

Ontario Works - Monthly statistical report sample
7.https://www.mcss.gov.on.ca/documents/en/mcss/social/reports/OW_EN_2017-12.pdf

Ministry of Children, Community and Social Services Annual report on Ontario Works

8.http://www.auditor.on.ca/en/content/annualreports/arreports/en18/v1_311en18.pdf

"Understanding Social Assistance In Northern Ontario: 1997 To 2010" by Anita Lynn Vaillancourt

9.https://tspace.library.utoronto.ca/bitstream/1807/77393/1/Vaillancourt_Anita_201611_PhD_thesis.pdf

Canadian Research Data Centre Network – Open data initiative

10.https://crdcn.org/datasets/osad-ontario-social-assistance-database , Ontario Social Assistance Database

11.https://crdcn.org/publicationsearch??keys=&tid_i18n_3%5B%5D=4484, List of publications using the data

12.” Returning to Ontario Works” by Ernie Lightman and Andrew Mitchell

https://search.proquest.com/openview/7a716f813b172d9b20f66e6b2d43dcc0/1?pq-origsite=gscholar&cbl=28163

7.APPENDIX