Identifying Geographical Areas using Machine Learning for Enrolling Women in the Canadian Armed Forces

Ryuichi Ueno, Peter Boyd[a] and Dragos Calitoiu[b]
Department of National Defence, 60 Moodie Drive, Ottawa, ON, K1A 0K2, Canada

Keywords: Feature Selection, Clustering, Logistic Regression, Propensity Scores.

Abstract: To improve the visibility of military service as a career option for women, the Canadian Armed Forces (CAF) can tailor marketing campaigns to geographical areas and demographics within Canada that have historically high enrollment of women. To aid in this recruitment strategy, a logistic regression model was developed using historical recruiting data. The score obtained was used to rank Canadian postal codes and to identify the ones with the highest potential for recruiting of women. Additional demographic filtering was applied using marketing segments provided by a vendor. The final top 10% postal codes with the highest probability of women enrollment were clustered based on the collective social media behaviour of each postal code and was binned using the distance to the nearest recruiting centre. Several social media outlets were observed to be of interest, among them YouTube and Snapchat appear as viable options to reach women with a high probability of CAF enrollment.

1 INTRODUCTION

The Canadian Armed Forces (CAF) is committed to increasing the number of women from 15.9%[1] to 25.1%, as strategically emphasized in the latest defence policy (Strong, Secure, Engaged) (Department of National Defence, 2017). In order to rapidly increase the enrolment of women, tailored recruitment strategies must be used, focusing on attraction and selection of suitable women candidates for military service within specific military occupations. Efforts that are customized to different subgroups within a population based on past success can also contribute to increasing enrollment.

Ueno *et al.* applied unsupervised machine learning to group Canadian postal codes into clusters, using demographic information as features to partition the population (Ueno et al., 2019). The clusters were then ranked based on historical application and enrollment data. A ranking scheme was devised that considered both successful and unsuccessful applicants (men and women), making it adaptable to specific campaign requirements.

In the United States (U.S.), a similar approach of identifying geographical locations with potential for recruiting has been recently explored (Buttrey et al., 2018; Fulton, 2016; Monaghan, 2016). The U.S. Army uses commercial market segmentation data to analyze markets and past enrollment to assign recruiters and quotas to maximize production. Fulton (Fulton, 2016) proposed a clustering approach using ZIP codes, exploiting the different densities of potential recruits. Fulton's approach considered 347 variables from publicly available U.S. government agencies for all 34,007 ZIP codes to cluster them into similar groups. By using three dissimilarity measures and three clustering algorithms, he identified 18 clusters, "a new method of identifying potentially high production ZIP codes" and "an easier tool for recruiting commanders" (Fulton, 2016). Monaghan (Monaghan, 2016) expanded Fulton's work to the U.S. Navy by developing five-cluster membership factors based on publicly available data sources, for helping Navy Recruiting Command predict the number of leads as an indicator of the market depth that a ZIP code will produce.

The U.S. Army Recruiting Command (USAREC) uses a segmentation developed by Nielsen (former Claritas): Potential Rating Index for Zip Markets (PRIZM) (Nielsen Company, 2020). Each household is assigned to one of the 68 segments, describing demographic and consumer behaviour. Since 2007, USAREC augmented the PRIZM segments with atti-

[a] https://orcid.org/0000-0001-5023-1200
[b] https://orcid.org/0000-0003-0173-9846

[1] As of the end of 2019.

Ueno, R., Boyd, P. and Calitoiu, D.
Identifying Geographical Areas using Machine Learning for Enrolling Women in the Canadian Armed Forces.
DOI: 10.5220/0010186703070316
In *Proceedings of the 10th International Conference on Operations Research and Enterprise Systems (ICORES 2021)*, pages 307-316
ISBN: 978-989-758-485-5

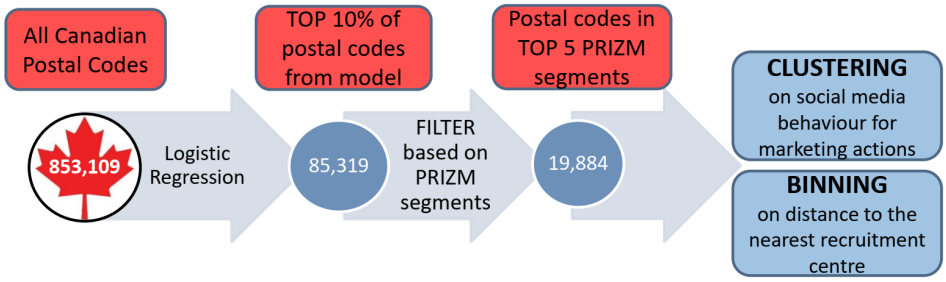

Figure 1: Overall flow of the method presented in this paper. Each of the steps are detailed in the respective sections in the text.

tudes, motivators and barrier to service data and developed 39 Army Custom Segments (Clingan, 2007), called tactical segments. USAREC also grouped them into 10 strategic segments and determined which segments have a population with a higher propensity for joining the military. Recruiters used this information to better decide who they should attempt to recruit and how.

PRIZM has not only been used for the purposes of military recruiting, these demographic categories are used predominately in the business industry for targeted marketing. Aside from these sectors, they are also used in a wide variety of fields to identify geographical and demographical dimensions of a particular attribute. For example, in retailing and manufacturing, PRIZM segments were used to predict in which geographical regions product affinity would be highest (Ayzenshtat et al., 2019). In urban planning, demographics of past and present inhabitants were found to dictate trends in urban landscape and vegetation characteristics (Boone et al., 2010). PRIZM segments were used also to identify sensitive demographics that live near hazardous waste sites in the US (Heitgerd, 2001).

The goal of this current research is to help recruiters identify geographical areas where women are more likely to join the CAF Regular Forces (RegF). In order to do this, we first used historical data, namely the home address of the women who applied to be hired as RegF members between January 2016 and March 2019. For each applicant, their postal code and a binary recruiting success variable (enrolled/not enrolled) were identified. A logistic regression model was built to discriminate postal codes for recruiting success. As independent variables, we explored 760 attributes that combine econometric, demographic and geographic characteristics at the postal code level. The logistic regression model produced a score that was applied to all 853,109 Canadian postal codes. Using the score as a ranking measure, the top 10% of postal codes were selected as the geographical locations of the women most likely to join the CAF.

The Canadian version of Nielsen's PRIZM segments were used to identify the locations producing women enrollments to complement the statistical model developed in-house. This PRIZM dataset contains 68 segments that capture demographics, lifestyle, consumer behaviour, and settlement patterns in Canada. The five most productive segments, covering 14% of the Canadian population, were able to collect almost 25% of all women enrollments.

The selected postal codes were further segmented to tailor to future recruiting campaigns. First, we used collective social media behaviour of each postal code to segment them and identify the best marketing approaches. Postal codes with similar social media behaviour were clustered using unsupervised machine learning. Second, the distances from each postal code to the nearest recruiting centre were computed, with the knowledge that the cost and efficiency of the recruiting process is correlated with the distance from the selected postal code to the recruiting centre.

The novelty of our proposal resides in the design of the process of searching for the best geographical location. The process contains two steps: (i) the selection, and (ii) the description of the selected location using two measures. The selection is done by combining an individual postal code score (obtained with logistic regression) with a collective score (from PRIZM segments [2]). The selection has these two separate scores because the PRIZM numbers are represented as categorical data and are not eligible to be independent variables in the logistic regression model. With these selections, we can help marketers with the description of the selection to identify the best communication channel and message to the potential applicants. Two measures are used for describing the selection: the social media behaviour and the distance to the nearest recruiting centre. We do not eliminate

[2]Compared to the collective score applied in the U.S., the Canadian PRIZM segments are derived at the postal code level. This granularity allows for a better modeling; the U.S. ZIP code is at least 10 times larger by population count compared to the Canadian postal code.

more based on our description of the selection, but instead we offer the marketers options to adapt their action plan to the proposed selection.

Section 2 details the datasets used for this work. Section 3 describes the logistic regression model. Section 4 lays out the outcomes obtained by applying the PRIZM segments. Section 5 introduces the clusters for customizing the marketing message. Section 6 presents the binning considering the distance from the home postal code of the applicant to the closest recruiting centre, followed by conclusions in section 7. See Figure 1 for the visual depiction of the overall methodology.

2 DATASET

Four datasets were used for this paper, all of which were obtained at the postal code granularity. Each of the datasets is described below.

2.1 Women Applicant Dataset

This dataset was retrieved from the Canadian Forces Recruiting Group (CFRG) and it contains women applicants to the CAF from January 2016 to March 2019. We removed applicants with no final decision, keeping only the enrollments and non-enrollments. We retrieved postal codes from each applicant and they were counted and labelled into 1,513 postal codes with enrollments and 12,507 postal codes with non-enrollments. If a postal code had both enrollments and non-enrollments, it was labelled as as an enrollment.

2.2 DemoStats

A demographic dataset called DemoStats is provided by Environics Analytics, a data, analytics and marketing services company specializing in geo-demographic segmentation (Environics Analytics, 2018). A total of 760 demographic attributes were considered as potential candidates to explain the target variable, the enrollment. DemoStats is built from a variety of data sources, including the latest census, economic indicators, postcensal estimates from the federal and provincial governments, immigration statistics and economic data. It features variables on population, family structure, household size and type, ethnic diversity, labour force participation and income.

2.3 PRIZM Segments

The PRIZM segments used for this paper are the Canadian version of Nielsen's PRIZM segments as described in the previous section. This dataset was also acquired from Environics Analytics, and is used to identify the locations producing women recruits. The PRIZM dataset contains 68 segments capturing demographics, lifestyle, consumer behaviour and settlement patterns across Canada. Since this dataset is entirely categorical, it is used as a filtering tool, not as a dataset for modelling.

2.4 Opticks Social

The social media dataset consists of 720 social media attributes in Opticks Social provided by Environics Analytics. These attributes are derived from an online survey of 19,938 respondents from AskingCanadians^{TM}, and describes social media usage, frequency and behaviours. The survey data are matched to postal code by using the k-nearest neighbour approach in conjunction with 50 Demostats attributes and match the postal code with the smallest differences.

3 LOGISTIC REGRESSION MODEL (DEVELOPING AND SCORING)

We have compared logistic regression model with two neural network models (a multi-layer perceptron, MLP, and a deep neural network, DNN) in a related study on a dataset involving all recruits, and found that the difference in performance was marginal for previous dataset. The much smaller size of the training set used for this study (since it focuses on women) makes it unsuitable for training neural networks and, therefore, we opt to use logistic regression for scoring the postal codes.

Logistic regression models the probability of belonging to a *class A* when there are only two mutually exclusive classes, namely *class A* (enrollment) and *class not-A* (non-enrollment). Instead of using a linear regression model to represent these probabilities:

$$p(X_1, X_2, ..., X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \quad (1)$$

where $X_1, X_2, ..., X_n$ are the dependent variables, we modelled a function that gives outputs between 0 and 1 for all values of X_i. Logistic regression does this by using the logistic function:

Table 1: The most significant features found from logistic regression.

Name of the Variable	Contribution to the model	Pr>ChiSq	Parameter Estimation	Mean	Min	Max	Std	Cap	VIF
intercept			-2.36720						
ECYINDPUBLP	74.81%	<0.0001	0.00594	10.93	0	1366	39.43	129.22	1.00
ECYACTUR	9.84%	<0.0001	-0.00206	7.19	0	62.5	6.48	26.63	1.01
ECYINDWHOL	6.86%	0.0043	0.00677	5.32	0	600	17.52	57.89	1.18
ECYACTPR	5.77%	0.0008	0.00615	66.84	0	100	15.47	N/A	1.00
ECYOCCSCND	2.73%	0.0016	0.00394	8.03	0	1209	29.95	97.88	1.20

$$p(X_1, X_2, ..., X_n) = \frac{\exp[\beta_0 + \beta_1 X_1 + ... + \beta_n X_n]}{1 + \exp[\beta_0 + \beta_1 X_1 + ... + \beta_n X_n]} \quad (2)$$

The logistic function always produces an S-shaped curve, and regardless of the value of X_i, we obtain a sensible prediction.

For deriving the logistic regression model, Base SAS version 9.4 was used. For the model development, the women applicant dataset was combined with the Demostats dataset using the postal code as joining key. The dataset was randomly split 50%-50% into development and validation sets. All Demostats variables are numerical, positive real numbers. A capping on the right with 3σ was implemented, a standard procedure for assuring the stability of the model. The cap values for the final selection are presented in Table 1. A combination of stepwise regression and multicollinearity was used to reduce the number of variables. SAS PROC LOGISTIC (EVENT='0') was implemented for the stepwise selection, with a significance level of 0.1 (SLE=0.1) required to allow a variable into the model, and a significance level of 0.01 (SLS=0.01) required for a variable to stay in the model. The Hosmer and Lemeshow goodness-of-fit test (Harnett, 1975) for the final selected model was requested by specifying the LACKFIT option.

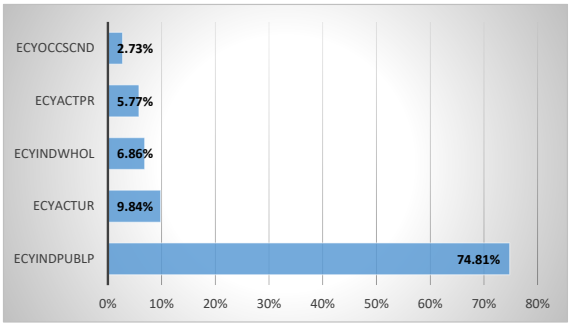

Figure 2: Variable contributions.

The contributions of the selected variables are presented in Figure 2. We report here only the parameters associated to the variables that have a significant contribution to the logistic regression model: ECYINDPUBLP (Household Population 15 years or over by Industry: Public Administration), ECYINDWHOL (Household Population 15 years or over by Industry: Wholesale Trade), ECYOCCSCND (Household Population 15 years or over by Industry: Occupations unique to manufacture and utilities), ECYACTUR (Household Population 15 years or over by Labour Force Activity: Participation Rate), and ECYACTPR (Household Population 15 years or over by Labour Force Activity: Unemployment rate). The participation rate represents the percentage of people who are in the labour force (namely employed or actively seeking work); unemployment rate represents the percentage within the labour force that is currently without a job.

The parameter estimations, namely the coefficients of the logistic regression for the target variable being zero, are presented in Table 1. Two descriptors regarding the possible multicollinearity (VIF[3] and Condition index[4]) are computed. SAS PROC REG /vif and PROC REG /collin were applied for removing collinearity. VIF is presented in Table 1. The Condition index was smaller than 10. Each record receives a score, namely *enrollment_score*, and we apply Equation 2 to map the score into a probability as written below:

$$p\_0\_1 = 1 - \frac{\exp[enrollment\_score]}{1 + \exp[enrollment\_score]} \quad (3)$$

[3]The variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least square regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. A rule of thumb is that if VIF > 3, then multicollinearity is high.

[4]The condition index is the standard measure of ill-conditioning in a matrix. It is computed by finding the square root of the maximum eigenvalue divided by the minimum eigenvalue. If the condition index is above 30, the regression may have a significant multicollinearity. Multicollinearity exists if two or more of the values related to the high condition index have a high proportion of variance explained. One advantage of this method is that it shows which variables are collinear.

Table 2: Logistic regression results.

Decile	Development data				Validation data			
	p_0_1	Enrollment		Enrollment	p_0_1	Enrollment		Enrollment
	Mean	Sum	ColPctSum	per decile	Mean	Sum	ColPctSum	per decile
1	0.1833	159	20.73	22.65%	0.1842	147	19.73	21.00%
2	0.1319	89	11.58	12.68%	0.1326	106	14.22	15.14%
3	0.1238	81	10.54	11.54%	0.1241	82	11.00	11.71%
4	0.1176	81	10.54	11.54%	0.1177	74	9.93	10.57%
5	0.1117	75	9.76	10.68%	0.1116	71	9.53	10.14%
6	0.1055	74	9.63	10.54%	0.1056	63	8.45	9.00%
7	0.0990	59	7.68	8.40%	0.0986	62	8.32	8.86%
8	0.0904	54	7.03	7.69%	0.0897	51	6.84	7.29%
9	0.0792	53	6.90	7.55%	0.0785	48	6.44	6.86%
10	0.0519	43	5.59	6.13%	0.0510	41	5.50	5.86%

We scored both the development and the validation set, which consist of 702 and 700 postal codes, respectively. A probability was computed from the score, and then the postal codes were sorted in descending order by the probability. We built deciles, ranking each postal code by the probability to enroll women and ordering from highest to lowest. We reported for each decile the following numbers: the size of the decile (number of postal codes), the *mean* of the enrollment probability predicted by the model (*p*), the number of enrollments realized on that decile (*sum*), the percentage of the enrollments collected in that decile out of total enrollments (*ColPctSum*) and, finally, the enrollment rate per decile. Table 2 contains these values for the development data and for the validation data. A model is powerful if it can group a high percentage of all enrollments in the first decile. For a random selection, we expect to collect 10% of all enrollments in the first decile. In our model, 20% of all enrollments were collected in the first decile, obtaining a lift of 200% (model vs. random). If the model has enrollment per decile descending from decile 1 to 10 without reversals (the stair case effect) then the model is statistically correct. Table 2 supports the lack of reversals in both the development and validation data.

In order to quantify the efficiency of the model, the Kolmogorov-Smirnoff (KS) statistic is computed[5]. For the development data set, KS=0.1559; for the validation data set, KS=0.1346, both of them being greater than 0.1, an acceptable value.

We extend this model to score the entire 853,109 Canadian postal codes. A probability is computed from the score, and then the postal codes are sorted in descending order by the probability. We select the top 10% for the next steps (85,310 postal codes).

4 FILTERING WITH THE PRIZM SEGMENTS

Using all of the women applicant data and all of the postal codes in Canada, it is found that the five most productive PRIZM segments are able to collect almost 25% from all enrollments: PRIZM 32 (Mini Van & Vin Rouge) 6.60%, PRIZM 37 (Trucks & Trades) 4.84%, PRIZM 16 (Pets & PCs) 4.23%, PRIZM 24 (Fresh Air Families) 4.23%, and PRIZM 63 (Lunch at Tim's) 3.75%.

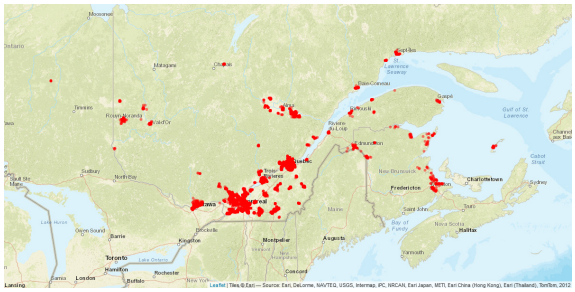

Figure 3: Geographical locations of the postal codes in PRIZM segment 32.

Out of 85,310 postal codes selected from the logistic regression, we identified 19,884 postal codes from the first five PRIZM segments. We ranked all the

[5]The Kolmogorov-Smirnoff statistic is used to measure the discriminatory power of a model, namely how the distribution of the score differs among enrollments and non-enrollments. The KS measures the minimum point of separation between the cumulative distribution functions of two distributions.

Table 3: Distribution of selected postal codes into PRIZM segments by age group.

PRIZM	Rank	postal codes	age 17–24	age 25–29
32	1	8,594	55,083	43,899
37	10	2,473	12,119	10,360
16	16	2,060	8,874	6,754
24	9	3,206	46,214	25,620
63	6	3,551	4,237	3,266

PRIZM segments in the first decile to validate where the five best PRIZM segments were distributed. The ranks are presented in Table 3. This additional filtering complements the statistical model and adds extra stability in the final selection. The distribution of these postal codes by PRIZM segment is presented in Table 3, together with the women population contained in these postal codes for three age groups: 17-19, 20-24, and 25-29. These statistics are useful since most of the recruits into the CAF are under age of 30 (Department of National Defence, 2018). The geographical distribution for PRIZM 32 is presented in Figure 3 (The distribution for all five PRIZM segments and their descriptinos are presented in the Appendix.

5 CLUSTERING ON SOCIAL MEDIA BEHAVIOUR

In order to assist marketing efforts for future recruits, we segmented the 19,884 postal codes into several clusters based on social media behaviours and made recommendations on how to best market to these populations. We employed an unsupervised machine learning technique to blind the actual recruit information, similar to the approach taken in (Ueno et al., 2019).

For this purpose, we used the Opticks Social dataset from Environics Analytics. This dataset was first standardized to remove means and to scale to unit variance, then aggregated down to 72 features using a Feature Agglomeration (FA) technique with Ward linkage criterion that minimizes the variance of the aggregated features (Ward, 1963). This method is implemented in the `Scikit-learn` library version 0.22.1 with Python version 3.7.1. (Pedregosa et al., 2011)

5.1 Developing Clusters

We used a Gaussian mixture model (GMM) to cluster the social media dataset. GMM assumes the data contains a finite number of clusters that are normally distributed, and uses an expectation-maximization (EM) technique to fit the mean and variance of the multivariate Gaussian distribution (Dempster et al., 1977). GMM assigns nodes to several clusters probabilistically in the absence of sharp separation boundaries. This technique is implemented in the `Scikit-learn` package.

One of the main and difficult problems in clustering is finding the right number of clusters to describe the observations. For solving this problem, we used the Bayesian Information Criterion (BIC) as defined by:

$$\mathrm{BIC} = ln(n)\,k - 2ln\left(\hat{L}\right) \tag{4}$$

where $\hat{L}$ is the maximized likelihood value, n is the number of data points, and k is the number of clusters. BIC will correct for over-fitting of a model by introducing a penalty for additional free parameters for higher k.

Figure 4 shows the BIC obtained as a function of the number of clusters, where the score for each cluster configuration is an average of 10 independent trials. As can be seen, the 7-cluster configuration is favourable. However, two of the small clusters were close in Euclidean distance; therefore they were merged together, and six final clusters were identified.

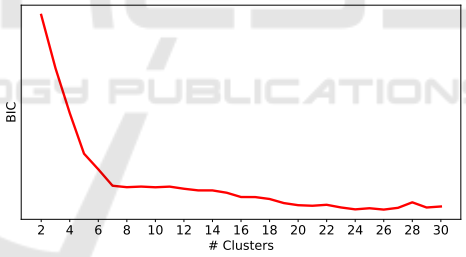

Figure 4: BIC scores of GMM clustering as a function of number of clusters. The BIC score is an average of 10 independent trials for each cluster configuration.

Table 4: Result of GMM clustering.

PRIZM	Cluster 1	2	3	4	5	6
32	7,776	37	0	23	746	12
37	2,020	9	0	0	444	0
16	1,235	25	0	0	800	0
24	2,398	70	28	0	710	0
63	3,474	4	0	0	73	0
Total	16,903	145	28	23	2,773	12

Table 4 shows the result of the GMM clustering, separated into the 5 PRIZM segments as presented before. As it can be seen, most of the postal codes be-

Table 5: Summary of social media (SM) behaviour by GMM cluster.

Cluster	Feature Importance	Summary of behaviours
1 & 5	12.1 %	- Current podcast listener, but not sports or technology focused podcast - Uses Wikis to read articles or for research a few times a week
2	12.3 %	- Typically uses mobile phone to access SM - Watches YouTube frequently, but not active on Instagram - Share content on SM forums and chats
3	11.3 %	- Snapchat user, but not active on LinkedIn - active on chats. Shares links with friends and colleagues
4	10.0 %	- Typically uses computer to access SM - Low use of YouTube, Instagram, or Facebook (a few times a month) - Uses SM for work/professional purposes than personal use
6	15.1 %	- Flickr user

long in cluster 1 and 5. Cluster 3 describes a small subset of PRIZM segment 24 while Clusters 4 and 6 describe a small subset of PRIZM segment 32. Figure 5 shows the distribution of the probability from the logistic regression model within each cluster. As can be seen, the clusters generally correlate with the probabilities from the logistic regression model and are able to distinguish the postal codes with strong probability from those with weaker probability.

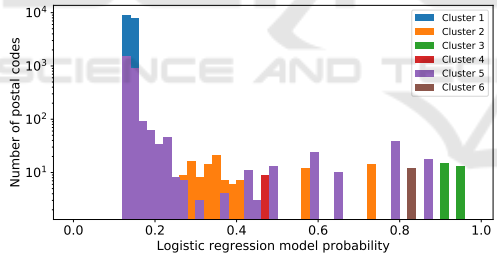

Figure 5: Distribution of postal codes by the logistic regression model probability. The colours correspond to the cluster of each postal code.

5.2 Social Media Behaviour of Clusters

In order to guide marketing strategies, we profiled each social media cluster identified above. In order to do so, we employed a random forest classifier (implemented in the `scikit-learn` library) to determine which features are most important for each of the six clusters. Since the social media attributes are aggregated into 72 groups of similar features, we were able to profile each cluster using several descriptions.

Table 5 shows the summary of the social media behaviour of clusters. As can be seen, the clusters portray distinct social media behaviour except for clusters 1 and 5, which are very close in terms of logistic regression probability (See Figure 5). For example, in order to reach the population in cluster 3, Snapchat is preferable to LinkedIn. For cluster 2, YouTube advertisement is a viable option for recruiting women.

Table 6: Histogram of distances from postal codes in the top 5 PRIZM categories to the nearest recruiting centre.

PRIZM	Distance (km) <25	25–50	50–75	75–100	>100
32	4,119	2,204	850	341	1,080
37	723	160	109	85	1,396
16	1,076	281	55	48	600
24	869	808	401	340	788
63	1,004	436	367	273	1,471
Total	7,791	3,889	1,782	1,087	5,335

6 BINNING BASED ON THE DISTANCE TO THE NEAREST RECRUITING CENTER

Interactions between recruiters and applicants vary based on the distance between an applicant's residence and the recruiting centre. Therefore, we computed the distance from each of the selected 19,884 postal codes to the closest recruiting centre. There are 26 recruiting centres in Canada. The distance was computed between the center of the postal codes of the recruiting centre and selected postal code described by longitude and latitude coordinates. We grouped the distance into five bins: less than 25 km, between 25 and 50 km, between 50 and 75 km, be-

tween 75 and 100 km, and more than 100 km. Almost 60% from all selected postal codes were less than 50 km from the nearest recruiting centre. The number of postal codes in each bin per PRIZM segment is presented in Table 6. This allows an additional tool for the marketers to focus on the recruiting strategy based on the distance.

7 CONCLUSIONS

In the absence of the individual data for the potential applicants, we explored the problem of identifying geographical areas with high potential for recruiting women by using demographic, life style, consumer behaviour, settlement patterns and social media characteristics of the neighbourhood (postal code) of the potential applicants. Using historical women recruiting data and two Environics datasets (DemoStats and PRIZM), a selection of 19,884 postal codes with highest recruiting potential was obtained from the most promising 10% of all Canadian postal codes, obtained using a logistic regression model. Finally, the selection was segmented to derive the optimum marketing channels and messages by using two approaches: (i) clustering based on the aggregated social media behaviour of the postal code and (ii) binning using the distance to the nearest recruiting centre. The social media behaviour was retrieved from a third Environics dataset, Optiks Social. The approach presented in this paper can be applied frequently for a long term, contributing to the commitment to increase women representation in CAF.

REFERENCES

Ayzenshtat, L., Rajamani, K., Vishnevskiy, A., Georgiev, N., and Preotescu, M. (2019). Methods and apparatus to identify affinity between segment attributes and product characteristics. US PATENT 0073685 A1.

Boone, C., Cadenasso, M., Grove, J., Schwarz, K., and Buckley, G. (2010). Landscape, vegetation characteristics, and group identity in an urban and suburban watershed: Why the 60s matter. *Urban Ecosystems*, 13(3):255–271.

Buttrey, S. E., Whitaker, L. R., and Alt, J. K. (2018). Developments in the statistical modeling of military recruiting. *CHANCE*, 31(2):38–44.

Clingan, M. (2007). U.s. army custom segmentation system. Presentation at 75th Military Operations Research Society Symposium, June 12-14, Annapolis, MD.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Department of National Defence (2017). Strong, secure, engaged: Canada's defence policy.

Department of National Defence (2018). Annual Report on Regular Force Personnel:2015-2016. Internal Reference Document: DRDC-RDDC-2018-D086.

Environics Analytics (2018). Demostats database. https://www.environicsanalytics.com/en-ca/data/demographic/demostats. Accessed Feb. 20, 2019.

Fulton, B. M. (2016). Determining Market Categorization of United States Zip Codes for Purposes of Army Recruiting. Master's thesis, Naval Postgraduate School, Monterey, CA.

Harnett, D. L. (1975). *Introduction to statistical methods*. Addison-Wesley Pub. Co.

Heitgerd, J. L. (2001). Using gis and demographics to characterize communities at risk: A model for atsdr. *Journal of Environmental health*, 64(5):21–30.

Monaghan, E. M. (2016). Estimating the depth of the navy recruiting market. Master's thesis, Naval Postgraduate School, Monterey, CA.

Nielsen Company (2020). My best segments. https://claritas360.claritas.com/mybestsegments/. Accessed Jan. 31, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ueno, R., Bryce, R., and Calitoiu, D. (2019). Ranking Clusters of Postal Codes to Improve Recruitment in the Canadian Armed Forces. In *Proceedings of the 18th IEEE International Conference on Machine Learning and Application*, volume 1, pages 1192 – 1197. Boca Raton, FL, USA.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

APPENDIX

Maps of PRIZM Segments

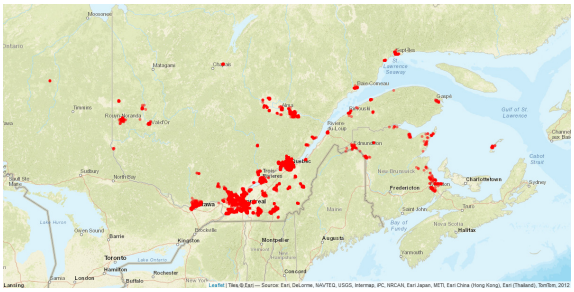

Figure 6: Geographical locations of the postal codes in PRIZM segment 32.

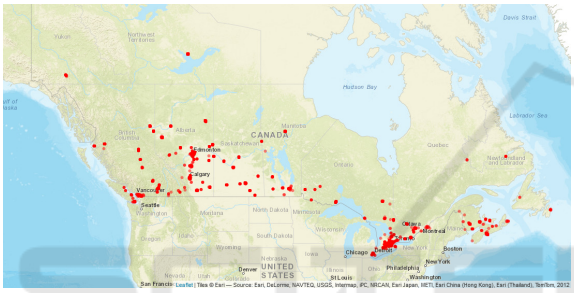

Figure 7: Geographical locations of the postal codes in PRIZM segment 37.

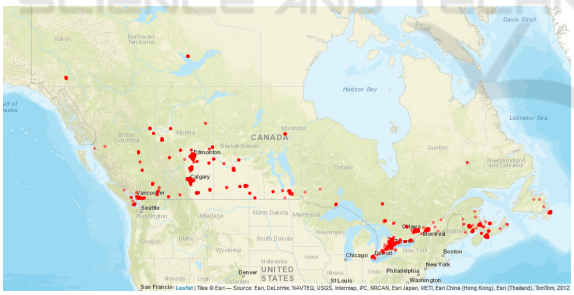

Figure 8: Geographical locations of the postal codes in PRIZM segment 16.

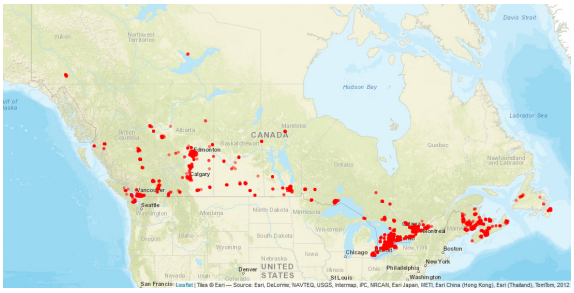

Figure 9: Geographical locations of the postal codes in PRIZM segment 24.

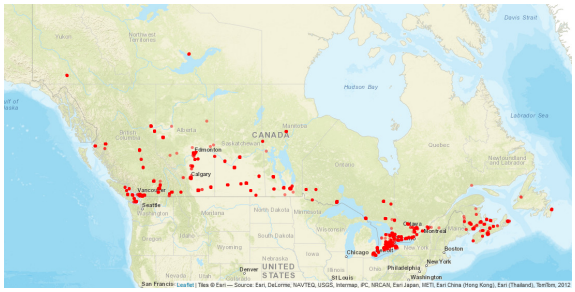

Figure 10: Geographical locations of the postal codes in PRIZM segment 63.

PRIZM Segment Descriptions

Below are the descriptions of PRIZM segments defined in the technical documentation provided by Environics Analytics.

PRIZM 32. Mini Van & Vin Rouge represents a collection of younger and middle-aged families and couples who live in new exurban communities beyond Quebec's big cities. These households consist of married and common-law couples. Although most households are French-speaking, more than 40 percent are bilingual. Their mixed educations provide good blue- and white-collar jobs in the construction and manufacturing sectors, as well as management roles within the public-service sector, resulting in upper-middle-incomes and active lifestyles. Residents here have gym memberships and enjoy sports like hockey, cross-country and downhill skiing. After all that fresh air and exercise, they reward themselves by picking up dinner at a chicken restaurant or kicking back with a glass of Shiraz in their single- or semi-detached homes. For a night out, they might head to the opera or a popular music concert; their idea of a vacation is anything from a resort package to a sightseeing tour around the U.S. and Caribbean.

PRIZM 37. Younger and middle-aged families comprise Trucks & Trades, where skilled tradespeople and blue-collar workers have built a comfortable lifestyle while accumulating tidy savings. Concentrated in Alberta and the Prairies, this segment has a disproportionate number of oil and gas workers who have sought out jobs in resource-rich lands over the past two decades. What workers may lack in education, they make up for with practical

skills in primary industries as well as the trades and transportation sector. Many families are younger and middle-aged – most children are under the age of 15 – and live in single- and semi-detached houses built between 1961 and 1990. There's also an above-average presence of mobile dwellings hauled in to accommodate the sudden influx of workers. When not working hard, these households play hard: fishing, hunting, golfing, ATVing, snowmobiling and playing baseball, along with other sports. They also have high rates for owning boats, camping trailers and motorcycles.

PRIZM 16. One of the largest lifestyles in Canada, Pets & PCs is a haven for younger families with preschool children in the new suburbs surrounding larger cities. Nearly half of the children in this segment are under the age of 10, while many of the maintainers are under 45. Pets & PCs households have a strong presence of immigrants from China, the Philippines and India. Few segments have more new housing than this group; most residents have settled into a mix of single-detached, semi-detached and row house developments. With upscale incomes, segment members have crafted an active, child-centred lifestyle. These families participate in many team sports, including baseball, basketball and hockey, and they shuttle kids and their gear to games in spacious SUVs–typically newer models. On weekends, they head to kid-friendly destinations like zoos, aquariums and amusement parks. They fill their homes with an array of computers and electronic gear, including video game systems, tablets and just about anything that will occupy their children while the moms and dads grab the occasional date night at the movies or dinner at their favourite seafood restaurants.

PRIZM 24. Widely dispersed across Canada, Fresh Air Families is one of the largest segments–and growing. Found in rapidly expanding exurban communities, these neighbourhoods feature a mix of middle-aged couples and families with children of a broad spectrum of ages. While most adults have high school, trade school or college educations, these dual-income households enjoy solid, upper-middle-income lifestyles thanks to positions in public administration, construction and the skilled trades. They own single-detached homes, typically built since 1990, and nine out of 10 commutes by car to jobs in nearby suburbs. With its couples and families, the segment scores high for a range of marketplace preferences, frequenting big-box retailers, large department stores and discount grocers. Members of Fresh Air Families enjoy the great outdoors, particularly fishing, boating, canoeing and camping. Indeed, some of their favourite leisure activities are evident in their driveways, typically cluttered with boats, campers or motorcycles – and pickup trucks to haul them to parks and campgrounds. But they also enjoy indoor pursuits like crafting and knitting.

PRIZM 63. Located in dense, industrial neighbourhoods scattered across mid-sized cities, Lunch at Tim's consists of singles, families and solo-parent households living in older single-detached homes, semis and duplexes. They're the kind of tight-knit communities where residents enjoy socializing at local eateries like Tim Hortons–as well as pizza places, burger joints and fish-and-chip restaurants. With an unusually mixed age profile–it's no longer the bimodal segment of the past–Lunch at Tim's residents have above-average rates for residents who are single, divorced, separated or widowed; nearly half the adults in these neighbourhoods are unattached. Despite the lower-middle incomes, roughly two-thirds of households own their homes, mostly built before 1980. Residents enjoy quieter pastimes and have high rates for knitting and woodworking, as well as outdoor activities like hiking and swimming. When the mood strikes, they might play a friendly game of curling or splurge on tickets to a dinner theatre, baseball game or boat or craft show.