Improving Edit and Imputation Strategies through Feature Selection

Andrew Stelmack

Statistics Canada
Statistical Integration Methods Division
100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6
andrew.stelmack@canada.ca

Abstract. The Canadian Census edits and imputes missing and erroneous data using a nearest neighbor donor imputation methodology. The choice of auxiliary features and their respective weights used in the calculation of the similarity measure for the nearest neighbor algorithm can have a large impact on the quality of said imputation strategy. In the past, this choice was mainly influenced by subject matter expertise. For the 2016 Census however, in particular for some questions related to immigration, it was decided that feature selection would be employed to aid in the choice of features. This paper will describe and evaluate the chosen method for the 2016 Census, the Relief algorithm, as well as test and compare it with other feature selection methods. The methods are tested using Monte Carlo simulation studies with data on immigration category, taken from the 2016 Census, under various response mechanisms.

Key Words: Imputation, Nearest Neighbor, Feature Selection, CANCEIS, Information Theory

The content of this paper represents the position of the author and may not necessarily represent that of Statistics Canada.

1. Introduction

The editing and imputation of survey data is an important and time consuming undertaking within the survey process. At Statistics Canada, imputation methods vary from survey to survey, however one that is commonly used is nearest neighbor donor imputation. This is the method employed by the Canadian Edit and Imputation System (CANCEIS), the system used to impute missing values from Canada's Census. Nearest neighbor donor imputation finds, for each record requiring imputation (failed record), the most similar records (nearest neighbors) among those not requiring imputation (passed records). Among these most similar records, a single record is chosen at random to be the donor record who "donates" its value for the feature(s) requiring imputation to the failed record. This differs slightly from the traditional nearest neighbor algorithm seen in many machine learning applications where the values of the nearest neighbors are either averaged (for numeric features) or the mode (for categorical features) is taken.

The quality of the imputation procedure depends largely on the calculation of the similarity measure used to determine the nearest neighbors. Rather than Euclidian distance which is typically used in the nearest neighbor algorithm, CANCEIS employs a distance metric that is equal to the weighted sum of penalty functions to calculate similarity.

$$D_{fp} = \sum_i w_i D_i(V_{fi}, V_{pi})$$

Where w_i is a weight assigned to feature i and $D_i(V_{fi}, V_{pi})$ is a penalty function on [0,1] determining similarity between records f and p for feature i. This distance metric is invariant to scale and allows the user to choose from a host of penalty functions designed for various feature types; survey data is often categorical, something not easily handled by traditional similarity metrics. The weight, w_i, allows the user the ability to assign a larger weight to features with more predictive ability of the variable requiring imputation. Thus, there are two main decisions to be made by the user with concerns to this similarity measure; which features to include in the similarity formula, and once the feature set is chosen, how to weight them?

This paper will look to address these questions specifically with regard to the weighting of the selected features. The first section will give a brief overview of a sample of selected feature weighting methods. The second section will describe a dataset from the 2016 Canadian Census which will be used to compare the previously described feature selection methods. The third section will describe a simulation study under which imputation is performed using weights obtained by the described feature selection methods. Results and conclusion will follow.

2. Feature Selection Methods

In the past, imputation strategies for the Canadian Census made the decision of which features to use and how to weight them based largely on subject matter expertise. While the resulting imputations went through many rounds of certification and are undoubtedly of high quality, the sometimes subjective nature of decisions made solely based on subject matter expertise can lead to potential losses in quality. For instance, it may be clear that both sex and geography are variables that are predictive of income and eye colour is not, but is geography more predictive than sex? How much more? Twice as much? Three times as much? These are decisions that are not always clear and could benefit from the help of data driven solutions. For the 2016 Census, it was decided that for one of the variables requiring imputation; a new topic for 2016 pertaining to the program under which the respondent immigrated to Canada, we would endeavour to use a data driven method in conjunction with subject matter expertise.

We will now briefly describe six feature selection methods that will be used in a simulation study later in this paper. Further expositions on each method can be found in their respective papers or in the paper by Robik-Sikonja (2003) which looks at a majority of the six algorithms. Each of the six methods described below is a filter type feature selection method. Filter methods are those that assign a measure of predictive ability to a feature based on a relationship between that feature and the variable of interest. For example, this could be as simple as the absolute correlation between the two. These measures can easily be translated into weights, and features below a certain weight can be removed from the model.

All of the methods will be implemented in R using the CORElearn package (Robnik-Sikonja and Savicky, 2018) except for the random forest method which will be implemented using the randomForest package (Liaw and Wiener, 2002).

2.1 Relief

The method that was used for the immigration topic described in the previous section was the Relief (Kira and Rendell, 1992) family of algorithms, specifically the ReliefF algorithm (Kononenko, 1994). ReliefF uses the same underlying assumptions that the nearest neighbor algorithm does; a feature that is good at predicting the variable of

interest should have similar values between records that have equal values in the variable of interest. For example, if income were a good predictor of sex, and a male had a high income level, then one would expect other males to also have high income levels and conversely females to have low income levels.

The method by which ReliefF employs this assumption is by taking a random record from the data and then finding that record's nearest neighbor from the same class for the variable of interest (nearest hit) and its nearest neighbor from all of the differing classes (nearest misses). It then updates a vector of weights positively, anytime the features differ between the random instance and the nearest misses, and negatively anytime the features differ between the random instance and the nearest hit. This process is repeated (ideally for each record in the dataset) or until the weight vector stabilizes. A more thorough explanation can be found in Kononenko's paper.

2.2 Information Gain

Information gain (IG) calculates the difference between the impurity of a variable of interest before and after conditioning on a feature. Information gain uses Shannon entropy (H) to measure impurity, however other impurity measures could also be used (as in the next method; DKM). Given a variable of interest Y with possible values $\{y_1 \ldots y_m\}$ and a feature X with possible values $\{x_1 \ldots x_n\}$:

$$IG = H(Y) - \sum_{j=1}^{n} \left(P(x_j) \left(H(Y|x_j) \right) \right)$$

Where $H(Y) = -\sum_{i=1}^{m} P(y_i) \log_2 P(y_i)$, $P(x_j)$ the probability of observing class j in x and $P(y_i)$ the probability of observing class i in y.

2.3 DKM (Dietterich, Kerns, and Mansour, 1996)

The DKM method, so named for the authors of the paper introducing it, as mentioned previously uses the same functional form as information gain however replaces $H(Y)$ in the calculation of IG with the following:

$$DKM(Y) = 2\sqrt{\max_{i=1,2,\ldots,m} P(y_i) \left(1 - \max_{i=1,2,\ldots,m} P(y_i) \right)}$$

2.4 Information Gain Ratio (Quinlan, 1986)

Information gain adjusted to correct for the bias of information gain towards features with high cardinality:

$$IGR = \frac{H(Y) - \sum_{j=1}^{n} \left(P(x_j) \left(H(Y|x_j) \right) \right)}{H(X)}$$

2.5 Random Forest (Breiman, 2001)

The random forest algorithm calculates a measure of feature importance known as the mean decrease in accuracy. After building the forest, the out of bag (OOB) error rate is calculated. Then, one at a time, each feature in the dataset is permuted and the OOB error rate is calculated once more. The resulting decrease in accuracy can be considered a measure of feature importance. That is, if after a feature is permuted there is only a small decrease in accuracy, that feature must not have been utilized often in the

forest, or used in less important splits. This method has previously been used by Statistics New Zealand (Zabala, 2015) to develop weights for use in CANCEIS.

2.6 Cost Sensitive ReliefF (Robik-Sikonja, 2003)

Feature selection methods typically assume that the cost associated with errors are constant across classes, however this is not always the case. Particular classes may be more difficult to predict due to data imbalance or a lack of separable data, or it may be the case that the user would like to put more emphasis on predicting certain classes well. For example, in predicting cancer diagnosis making false negatives may be more harmful than making false positives. To this end, we will look at a cost-sensitive version of the ReliefF algorithm, the "average cost ReliefF" algorithm. Which modifies the standard ReliefF algorithm to use information provided by the user in the form of a cost matrix. See Robik-Sikonja (2003) for details on the exact process.

3. Dataset

For the 2016 Census, two new variables were added; the immigration category under which the respondent immigrated to Canada (admission category) and applicant type. These variables were added through a record linkage between an administrative immigration file and a database of census responses. Once linked, the variables were processed and disseminated among the other census variables, allowing for connection of these admission characteristics to census variables from the questionnaire. Similar linkages between administrative data on immigration characteristics and the Census have occurred in the past, however this was to be the first time that the variables would be processed (edited and imputed) allowing for higher quality analyses.

Records in this dataset may be missing and require imputation for various reasons. Primarily, records that were immigrants according to their census response but were not able to be linked in the record linkage process will have missing values for the admission category variables. Secondly, any records that were linked but whose administrative information was not cohesive with their census response had the admission category value set to blank as the census response took precedence over the linkage result. For instance, there were cases where the linkage result implied the respondent immigrated under a category that did not exist during the year that the respondent had given as their year of immigration on their census form. Even though it was possible the respondent gave the incorrect year of immigration, the census responses as a rule were kept and the linkage result was removed.

The admission category variable contains 26 categories that a respondent could immigrate under, stemming from 4 broader categories; Economic Immigrant, Immigrant Sponsored by Family, Refugee, and Other Immigrant. The applicant type variable contains 3 categories; Principal Applicant, Spouse of a Principal Applicant, and Dependant of a Principal Applicant. For this paper, we will focus only on the imputation of the admission category variable and for the sake of simplicity only at the level of the 4 broader categories. The data used in this paper is with regard to a subset of the immigrant population that were relatively difficult to impute, specifically, immigrants who have a spouse that either is not an immigrant or that immigrated prior to 1980 (the admission category variable exists only for persons immigrating since 1980). Imputation within this group is challenging as the admission category variable is highly correlated among spouses and since one of the spouses is not an immigrant (or immigrated prior to 1980) this correlation does not exist with which to draw from, leaving data that is more difficult to separate.

4. Feature Selection Results

The feature selection methods described in Section 2 were performed on a subset of the data described in the previous section. This subset was comprised of 60,000 records. Each method was performed on 16 features provided by subject matter experts thought to have predictive ability of the variable of interest.

In an attempt to better impute the minority classes, for the cost sensitive relief method, we will use a cost matrix where each element, $[i, j]$, below the diagonal is equal to the number of records in class *j* divided by the number of records in class *i*:

$$CM = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0.7 & 0 & 1 & 1 \\ 4.1 & 6.3 & 0 & 1 \\ 51.0 & 77.5 & 12.4 & 0 \end{bmatrix}$$

Where row/column 1 represents the Economic Immigrant class, row/column 2 the Immigrant Sponsored by Family class, row/column 3 the Refugee class and row/column 4 the Other class. That is, in this sample of data the Economic Immigrant class is 51 times more prevalent than the Other class, and according to this cost matrix, incorrectly imputing a record from the Other class as Economic Immigrant costs 51 times as much as imputing a record from the Economic Immigrant class as Other.

The results of each feature selection method for this subset of data can be seen in Table 1 below. The weights obtained by each algorithm have been multiplied by 1000 and rounded to the nearest integer for readability.

Table 1. Weights Obtained for Features (x1000 and rounded)

	Relief	IG	IGR	Relief Cost	DKM	Random Forest
Place of Birth	149	275	47	220	73	69
Year of Immigration	74	35	37	131	19	36
Age at Immigration	70	108	131	113	36	73
Highest Education Level	25	32	10	23	13	17
Marital Status	21	42	42	17	15	26
Employment Income	21	23	96	3	14	15
# of Family Members	18	5	2	3	2	5
Province of Residence	18	5	2	18	2	5
Employment NAICS	17	30	8	3	16	11
Location of Study	16	67	43	37	26	22
Spouses Admission Cat.	15	19	28	14	13	9
Employment NOCS	11	34	14	0	17	10
Sex	9	1	1	4	0	3
Official Language	8	11	12	1	6	5
Low Income	3	3	6	-2	2	0
Quebec Resident	2	2	2	-2	1	2

The feature selection methods largely agree in terms of the relative importance of the features, with some exceptions. The place of birth, year of immigration, and age at immigration features all show high predictive ability of the variable of interest across all feature selection methods. This result is in line with subject matter expertise and their

expectations. On the other hand, subject matter experts had expected that variables related to geography and income would have had larger weights than they did.

Lastly, of note is the discrepancy between the information gain ratio (IGR) results and the other methods, particularly with regard to the place of birth feature and employment income. Among all but IGR the place of birth feature is by far the most important feature, whereas in IGR it is only the third most important feature and has a much lower weight. In IGR the employment income becomes the second most important feature with a very high weight whereas in the other methods it is usually seen as important but middle of the pack so to speak.

Table 2. Average Relative Rank of Weight

	Average Rank	Standard Deviation
Place of Birth	1.5	0.76
Age at Immigration	2.0	0.82
Year of Immigration	3.7	1.49
Location of Study	4.8	2.41
Marital Status	5.3	1.25
Highest Education Level	7.0	2.31
Employment Income	7.3	3.04
Employment NAICS	8.7	1.60
Employment NOCS	9.0	3.16
Spouses Admission Category	9.2	1.34
Province of Residence	11.0	3.06
Official Language	11.5	1.61
Number of Family Members	11.7	2.29
Sex	14.0	2.52
Low Income	14.3	1.25
Quebec Resident	15.0	1.00

Table 2 above shows the average relative ranking of each feature and the standard deviation of those ranks. For most features, the standard deviation is quite low, again implying that the feature selection methods generally agree for most features.

5. Simulation Study

5.1 Setup

Two further samples were taken from the data of Section 3 (distinct from the 60,000 records used for feature selection) with these two samples also being distinct from each other. The first sample consists of 30,000 linked records with valid values and will be used to conduct the simulation study. The second sample consists of 20,000 linked records with valid values and 20,000 records that were either unlinked or not cohesive with their census response. The second sample will be used to develop a response propensity model that will be used to assign missing at random (MAR) response probabilities to the first sample.

For each simulation, 1000 replicate datasets were created of the first sample described above, with each record in each replicate dataset responding or not according to a Bernoulli random variable using an assigned response probability. Those records

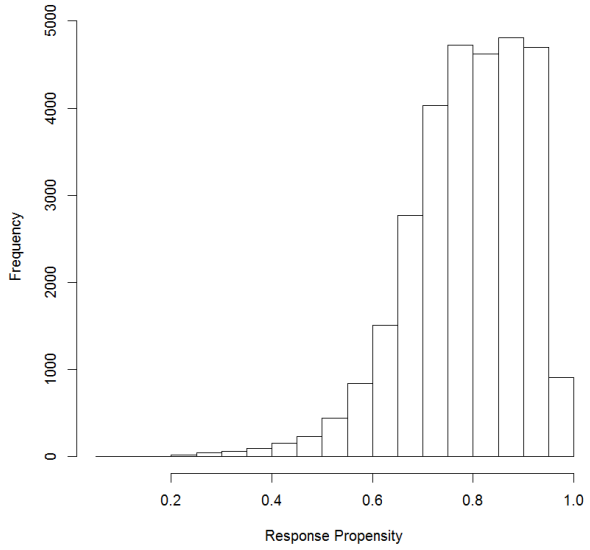
receiving a value of 0 for this random variable are said to not respond and have their value for the admission category variable set to blank to be imputed in CANCEIS.

Each replicate dataset is then run through CANCEIS and the "non-respondents" have their value for the admission category variable imputed using the weighting schemes of Table 1. Features with a weight under 10 in Table 1 were excluded for that method. After imputation is complete, the imputed values are compared to the "true" values (the values the records had prior to being set to blank) and Monte Carlo measures of relative bias (RB), relative root mean squared error (RRMSE), and the coefficient of variation (CV) are calculated for the estimate of the population proportion for each category. As it is important to not only have quality estimates at the population level but also at the record level, analysis is also completed at the record level with measures of accuracy, precision, recall, and F1 score (the harmonic mean of precision and recall) calculated.

5.2 Response Propensities

The simulation study will be performed under three response mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Under the MCAR mechanism, each record will be assigned an 80% probability of response. To simulate the MAR mechanism, a response propensity model was created using lasso logistic regression (Tibshirani, 1996) in the glmnet package of R (Friedman et al., 2010) and the second sample described previously. The model was created using the same features suggested by subject matter for use in the imputation algorithm. Using the model, response propensities were predicted for the 30,000 linked records that were to be used in the simulation study. These records had an average probability of response of 78.78% with a distribution shown in the histogram of Figure 1. The minimum probability of response was 7.93% and the maximum was 98.63%. Under MNAR the response probabilities will be assigned as follows: Economic Immigrant – 90%, Immigrant Sponsored by Family – 70%, Refugee– 85%, and Other Immigrant – 80% for an average probability of response of 78.56%. These response probabilities were chosen after consulting a with subject matter expert who posited that the Immigrant Sponsored by Family class should have the lowest probability of being linked due to difficulty with regards to changes in family name over time.

Figure 1. Response Propensities for MAR Response Mechanism



5.3 Results

The results of the simulation study can be found in Figures 2 through 6 in the Appendix. Although all of the measures mentioned in the setup section were calculated, for the sake of brevity, these figures show only the composite measures; F1 score and RRMSE for all classes under the three response mechanisms. Also, in addition to the six feature selection methods, the simulations were run for two "baseline" measures; random imputation and an equally weighted imputation scheme.

As can be seen in Table 3 below, and Figures 2 through 6, throughout the study, for all of the response mechanisms, Information Gain, DKM, ReliefF, Cost Sensitive ReliefF, and Random Forest performed comparably for all measures, across all categories of the variable of interest, with ReliefF (and Cost Sensitive ReliefF) performing slightly better in most cases. As expected, they all outperformed the equally weighted scheme, which in turn outperformed random imputation. For these reasons, most of the comparisons drawn in this section will only be done between the equally weighted scheme and ReliefF and only concerning F1 Score and RRMSE.

Table 3. Macro Averaged F1 Scores

	MCAR	MAR	MNAR
Random	0.25	0.25	0.23
Equal Weight	0.38	0.38	0.35
IGR	0.40	0.41	0.37
IG	0.46	0.47	0.42
DKM	0.46	0.46	0.42
Random Forest	0.46	0.47	0.42
ReliefF	0.48	0.48	0.43
Cost Sensitive ReliefF	0.48	0.48	0.43

At the record level, Table 3 shows the macro averaged F1 scores across all four categories. In view of this, ReliefF outperforms equal weighting by 10 percentage points for each of the MCAR and MAR mechanisms and 8 percentage points for the MNAR mechanism. As the figures in the Appendix show, most of this increase can be attributed to the improvement in imputing the Refugee class. For that class, ReliefF outperformed equal weighting by 24, 25, and 21 percentage points under the MCAR, MAR, and MNAR response mechanisms respectively.

With respect to the cost sensitive version of ReliefF, it appears that the F1 score is behaving as one would hope. Tables 4a and 4b show the difference in F1 score between the standard ReliefF algorithm and the cost sensitive ReliefF version. As expected, the cost sensitive version improves the F1 score for the Refugee class as well as the Other class at the expense of the Economic Immigrant and Immigrants Sponsored by Family classes. The improvements in Table 4a appear small but looked at as percentages as in Table 4b, the percentage increase for the Other class is quite large.

Table 4a. Difference in F1 Score between ReliefF and Cost Sensitive ReliefF

	Economic	Family	Refugee	Other
MCAR	-0.0138	-0.0091	0.0053	0.0180
MAR	-0.0098	-0.0102	0.0029	0.0183
MNAR	-0.0135	-0.0086	0.0022	0.0125

Table 4b. Percentage Difference in F1 Score between ReliefF and Cost Sensitive ReliefF

	Economic	Family	Refugee	Other
MCAR	-2.31%	-1.25%	1.00%	37.26%
MAR	-1.63%	-1.44%	0.53%	29.48%
MNAR	-3.12%	-1.12%	0.45%	28.40%

With regard to the population proportion estimates and the relative root mean squared error, again, the ReliefF algorithm shows large improvements over the equal weighting strategy. Under each response mechanism, for most classes, the RRMSE was lower for ReliefF than for equal weighting. Table 5 shows the increase in RRMSE for equal weighting over ReliefF. The one exception where equal weighting performs better than Relief in terms of RRMSE is for the Refugee class under the MNAR response mechanism where it sees a small improvement. The other three classes however are worse off and of course at the micro level the F1 score is still worse for all classes. Therefore it is not suggestive of equal weighting outperforming ReliefF under the MNAR mechanism.

Table 5. Increase in RRMSE for Equal Weighting Compared to ReliefF

	Economic	Family	Refugee	Other
MCAR	0.32%	0.04%	2.12%	0.77%
MAR	0.09%	0.55%	4.57%	1.16%
MNAR	1.68%	0.90%	-0.97%	0.35%

It may be argued that an equal weighting scheme is not a strong baseline, as it could be expected that subject matter experts would at least be able to provide the most predictive variables and weight those accordingly and see improvement over equal weighting. While this is true, one could point to the one “outlier” selection method in this study; Information Gain Ratio. As previously mentioned, the IGR method scored employment income inordinately high and place of birth inordinately low. This likely caused the decrease in the quality of its weighting strategy that can be seen in view of the Figures 2 through 6 where IGR showed only slight improvements over the equal weighting scheme. While it is possible that subject matter experts would be able to provide a reasonable weighting scheme comparable to the four similarly performing selection methods, it is also possible that subject matter knowledge provide a weighting scheme comparable to IGR, which appears reasonable but performs worse. That is, imputation quality exists on a spectrum between the baseline and the optimum weighting scheme, and it is feasible to expect weights designed using subject matter knowledge to exist at any point on this spectrum.

6. Conclusion

All feature selection methods showed benefits to the quality of the imputation strategy for the tested dataset under various response mechanisms. Improvements were evident at the record level where gains of upwards of 10 percentage points over an equally weighted scheme were seen for the macro averaged F1 score. Gains were particularly strong for the Refugee class which saw improvements in F1 score of upwards of 25 percentage points.

While the relative ranking of features were similar across most of the feature selection methods, there were exceptions. And while the ReliefF algorithm outperformed the other methods nearly across the board, the differences were small. We also showed the relatively poor performance of the IGR method for this data when compared with the other methods. With this in mind, rather than declare that one method should always be used as a result of this study, it would instead be wise to test multiple methods, and along with subject matter expertise make decisions on weights in light of those tests. That is, feature selection methods should not necessarily be used with the intent of replacing subject matter knowledge, but rather used in conjunction with it, to allow for the most informed decision possible.

References

Bankier, M. (2011). NIM Technical Report, Unpublished report, Ottawa: Statistics Canada.

Breiman, L. (2001). Random Forests, *Machine Learning,* 45:5-32.

Dietterich, T.G., Kerns, M., and Mansour Y. Applying the Weak Learning Framework to Understand and Improve C4.5. *Machine Learning: Proceedings of the Thirteenth International Conference (ICML'96),* 96-103.

Freidman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate descent. *Journal of Statistical Software*, 33(1):1-22. URL http://www.jstatsoft.org/v33/i01/.

Haziza, D. (2003). The Generalized Simulation System (GENESIS): A Pedagogical and Methodological Tool. *2003 Joint Statistical Meetings – Section on Survey Research Methods.*

Kira, K. and Rendell, L.A. (1992). A Practical Approach to Feature Selection. *Machine Learning: Proceedings of International Conference (ICML 92),* 249-256.

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of Relief. *Machine Learning: ECML-94*, 171-182.

Kononenko, I. (1995). On biases in estimating multi-valued attributes. *Proceedings of the International Joint Conference on Artificial Intelligence*, 1034-1040.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2 (3):18-22.

Mcleish, S. (2017). 2016 Census of Population of Canada: Integration of Immigration Administrative Data. *UNECE Work Session on Migration Statistics.*

Quinlan, J.R. (1986). Induction of Decision Trees. *Machine Learning* 1:81-106.

Robik-Sikonja, M. (2003). Experiments with Cost-sensitive Feature Evaluation. *Machine Learning, Proceedings of ECML 2003*, 325-336.

Robik-Sikonja, M. and Savicky, P. (2018). COREIearn: Classification, Regression and Feature Evaluation. *R package version 1.52.1.* https://CRAN.R-project.org/package=COREIearn.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (methodological)*, 58 (1):267-288.

Zabala, F. (2015). Let the Data Speak: Machine Learning Methods for Data Editing and Imputation. *United Nations Economic Commission for Europe – Work Session on Statistical Data Editing.*

Appendix

Figure 2. RRMSE for MCAR

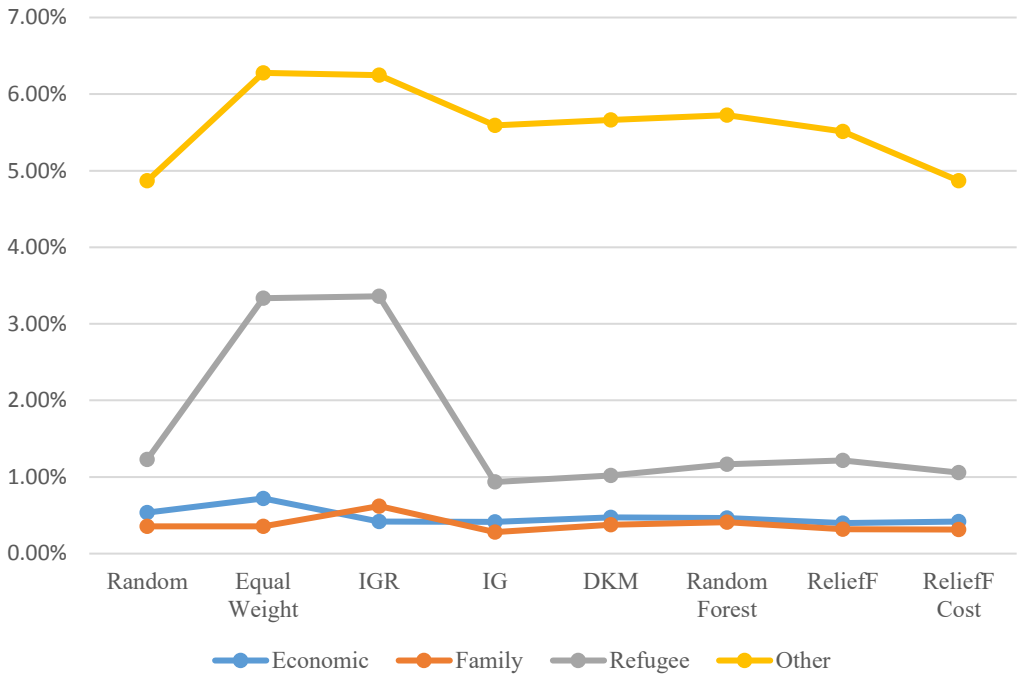


Figure 3. F1 Score for MCAR

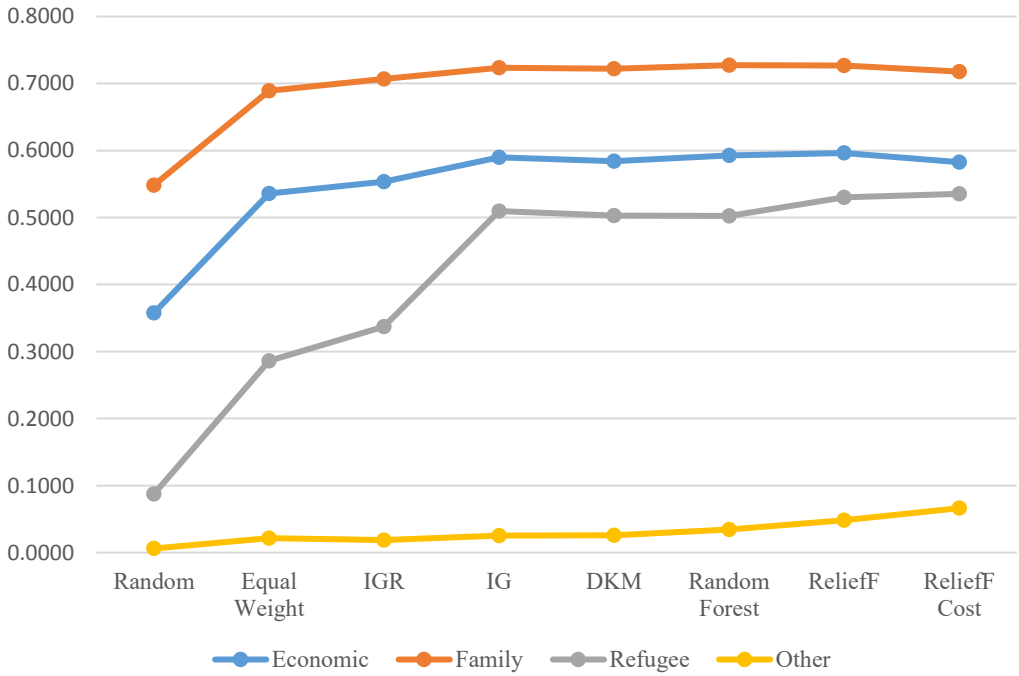


Figure 4. RRMSE for MAR

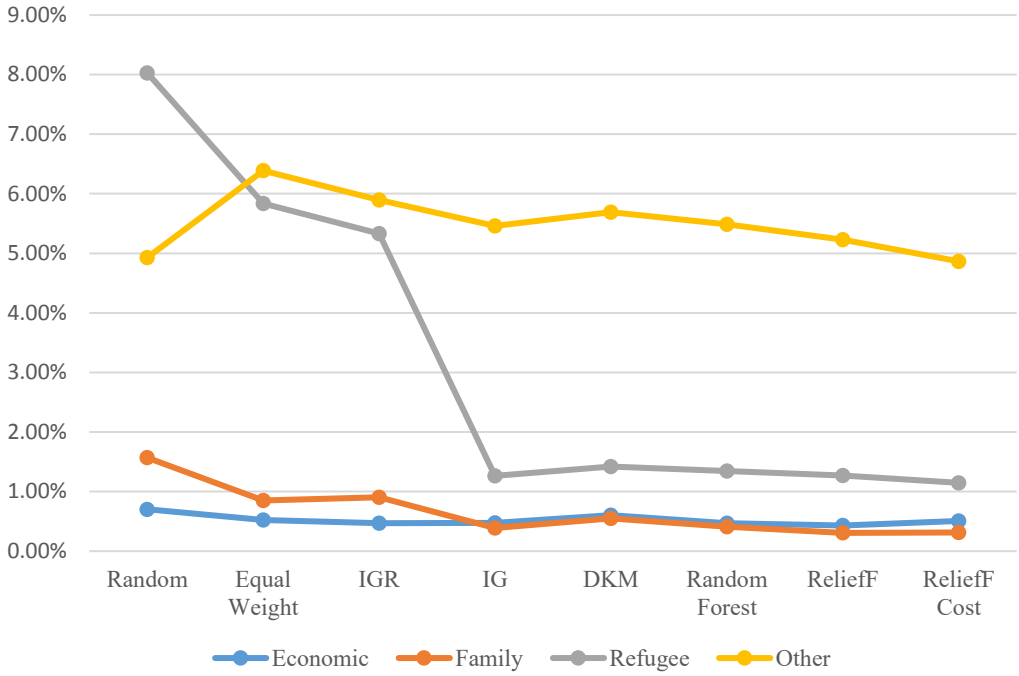


Figure 5. F1 Score for MAR

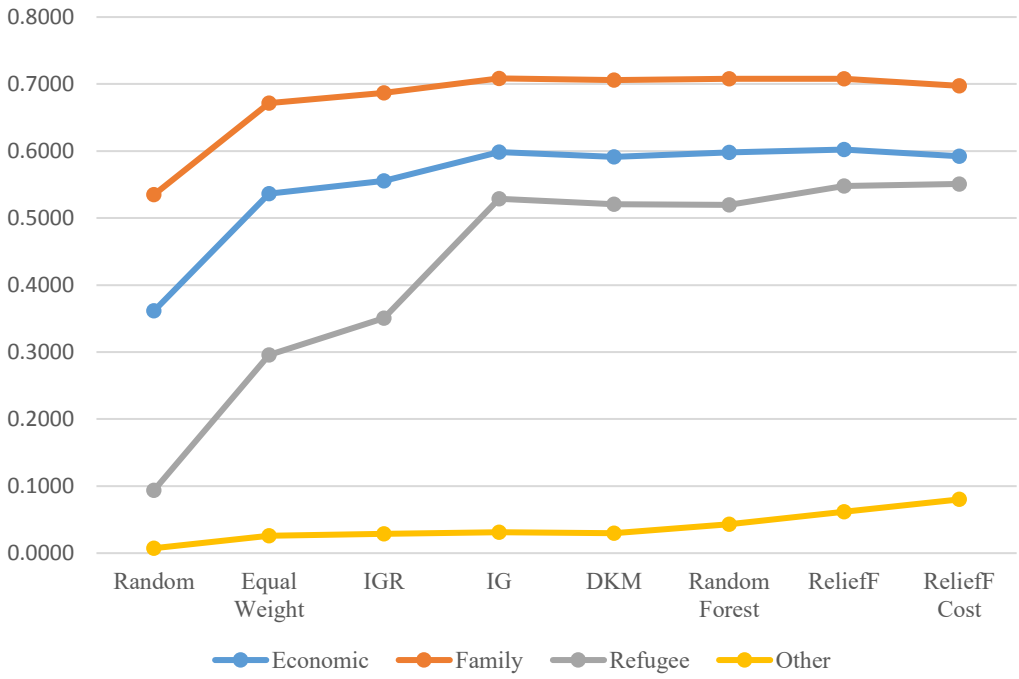


Figure 6. RRMSE for MNAR

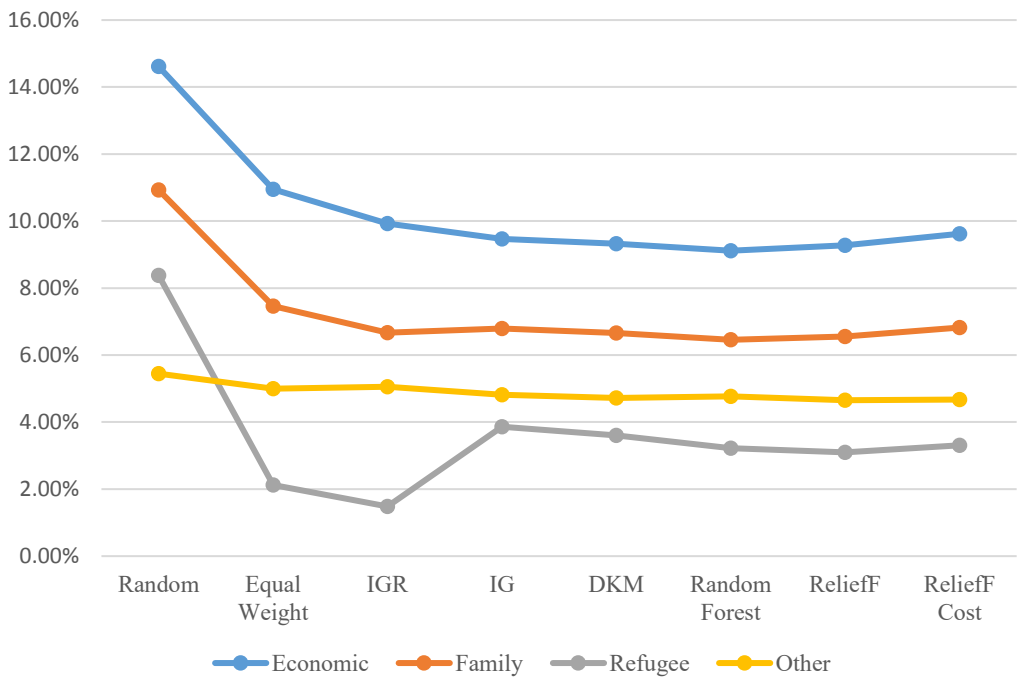


Figure 7. F1 Score for MNAR