KU LEUVEN

FACULTY OF SCIENCE

Predicting the Propensity to Move Using Public Register Data

Armen ABAGYAN
r0823936

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics and Data Science

Supervisor: Prof. Seppe vanden Broucke
Affiliation *KU Leuven*

Mentor: Michael Reusens
Affiliation *Statistics Flanders*

Academic year 2022-2023

Preface

The research presented for this thesis was conducted within the framework of a collaboration between KU Leuven and Statistics Flanders. The goal of the research was to develop a predictive model for moving propensity using Belgian public register data and machine learning algorithms. The wider implications for the study relate to the increasing use of secondary sources in research, and the shift away from primary sources.

I would like to personally extend my gratitude to my supervisor Professor vanden Broucke and my mentor Michael Reusens, who both provided me with the necessary guidance without which this thesis would not have been possible. I am also deeply indebted to many at Statistics Flanders who provided immeasurable support, direction, and encouragement throughout the more challenging moments of my work. These include, but certainly not limited to: Jan Pickery, Ingrid Schockaert, Lisa Van Landschoot, Ahmed Abdelhakim, and Noëmi Debacker. Lastly, I'd like to thank my partner Obay Al Bitar, for his enduring patience and loving support.

Summary

The present work endeavored to predict the propensity to move in Flanders using machine learning techniques and Belgian public register data, drawing from the Dutch study “Replacing a survey questionnaire by predictive modeling using register data” published in 2018 by Statistics Netherlands. To achieve this, several Belgian national registers were linked containing life history events and personal, household, neighborhood, and municipality level characteristics. Supervised binary classification methods elastic net penalized logistic regression, random forest, and XGBoost were employed for their superior performance with multicollinearity and high dimensional data. All features known up to a reference date of year t were used to predict moving behavior within 1 year of that reference date. January 1st, 2016, was used as a reference date to optimize and train the models, which were then applied to a test set with a reference date of January 1st, 2017. Hyperparameter tuning was performed for each of the methods via 10-fold cross validation on the training data, which was upsampled using the ROSE method to rectify class imbalance.

While prediction quality did not vary substantially with the method, a mean out of sample validation AUC of 81.5% (XGBoost) was achieved, indicating strong model discrimination between movers and stayers. The models generalize reasonably well to another year, though there is a drop in predictive performance (mean test AUC of 77.3% with the elastic net penalized logistic regression). When minimizing the False Positive Rate (FPR), or the proportion of incorrectly predicted movers, 52.17% of movers are predicted correctly and 83.35% of stayers are predicted correctly. When maximizing the True Positive Rate (TPR), or the proportion of correctly predicted movers, a TPR of 71.34% is achieved, though at the expense of an increased FPR (30.34%).

Though many improvements can be made to improve predictive accuracy, this study achieved the stated aim of predicting moving propensity using public register data. Such results corroborate the increasing viability of secondary data in lieu of survey questionnaires, and indicate that such methodology can be applied a wide variety of research questions

1. Introduction:

Survey Questionnaires have long been an essential tool in data collection and the backbone of research in the social sciences, allowing for the extraction of variables and constructs of interest. While methodological and technological innovations have ushered in more robust, representative sampling, time-saving and cost effective design, limitations persist in survey based approaches to data collection.

One such challenge plaguing surveys are non response, whereby the respondent does not respond to single questions (item non-response), or to the entire questionnaire (unit non-response). Even when a questionnaire is completed by the respondent, there exists the potential for a host of response biases which can have a negative impact on the accuracy and reliability of the collected data. Many such response biases, such as “Social Desirability” and “Acquiescence Bias”, are subject to psychological influences that inform the responses of survey participants and are difficult to control for in the survey design. [1]

Concurrent with the growing pressures on survey research due to lengthy processing time, non-response/ response bias, incomplete sampling frames, and high costs, is the increasing wealth of automatically generated non-survey administrative data.

Emerging from the automation of government services and digitalization of government records, administrative data maintains certain advantages over survey data. First, administrative records are often available for the entire population, allowing for larger sample sizes and more complete coverage. Second, administrative data is longitudinal in structure, permitting researchers to conduct individual level analysis over time. Third, it is rich in information often unavailable in survey sources due to non-response/response bias. Although not collected within the framework of a specific research design or question, secondary administrative data is increasingly being seen by researchers and statisticians as possessing viable research applications. [2,3]

The application of machine learning algorithms on secondary data has become more widespread in research across different disciplines, with many studies producing promising results. [4,5,6]

The present work draws on the 2018 study conducted by Statistics Netherlands ``Replacing a Survey Question by predictive modeling using register data” . In their work, the authors assessed whether the Dutch Housing Survey on someone’s desire to move houses could be replaced with a machine-learning predictive model based on public register data. The authors linked several registers from the Dutch System of Social Statistical Datasets (SSD) containing life history events from 1995-2016 and all features known up to a reference date were used to predict moving behavior within two years of that reference date. The features included time dependent and independent demographic features, time since last change in household composition, household, neighborhood and municipality characteristics, and whether or not someone is a homeowner or starter in the housing market. The study employed model based and machine learning techniques: logistic regression, lasso-regression, ridge regression, and random forest.

Both the ridge regression and random forest model identified age, time since latest change in household composition, and the time since the latest move or the number of moves over the last 17 years as the most important for predicting moving behavior. When optimizing for the difference between the True Positive Rate (TPR) and False Positive Rate (FPR) in selecting the cutoff threshold, the authors achieved a classification of 60% of movers correctly and 19% of stayers incorrectly. [7]

The ultimate goal of this research is to see whether such a study is possible in Flanders using Belgian public register data. It follows that the present study seeks to develop a predictive model for moving behavior within one year of the reference date given life history events, individual, household, neighborhood, and municipal characteristics known up to that reference date. Expanding upon the methodology used in the Dutch study, three machine learning binary classification techniques were employed: random forest, elastic net penalized logistic regression, and XGBoost.

2. Literature Review:

Socio-demographic characteristics such as age, gender and migration status have been well established as important predictors for moving behavior. Several studies have shown that younger people tend to be more mobile, opting for urban destinations, whereas older people become less mobile over time. [8,9]

Household composition and position is considered an important factor in one's moving propensity, with households without children are known to be more mobile than those without, since singles and couples without children are likely to be less settled. [10] Additionally, individuals living in a large household have been known to be more mobile. [11]

Family ties were shown to significantly diminish the mobility of individuals, whereby the existence of familial attachments in a municipality decreased the likelihood of out-mobility across ethnic groups. [12]

Since personal income indicates the extent to which housing is within financial reach, it likely impacts mobility behavior. Higher income individuals, therefore, may have greater number of dwelling options available, and may have greater ease of realizing an initial intention to move houses as compared to those with a low income status. Similarly, a higher level of education, often correlated with higher income, may also facilitate the realization of initial moving intention since more housing options are within financial reach. [13]

Employment status has been shown to have an impact on residential mobility. In a British study investigating the relationships between housing tenure, employment status and residential mobility, it was demonstrated that someone unemployed has a higher probability of moving than an employed individual with similar characteristics, though the propensity to move declines with unemployment duration. Additionally, the self-employed have a marginally higher probability of moving than the employed. [14]

Life history events are known to either trigger or attenuate mobility propensity, leading to the “postponement or cancellation of an initial intention to move'', but “may also constitute an unanticipated trigger for moving and subsequent move among people who had not intended to move before the event took place. Such events include “unanticipated” and “anticipated” life events, such as the formation and break up of unions, childbirth, a death in the household, and change in labor market status. While certain unanticipated events may trigger a move within a short time frame, anticipated events tend to take place after the move some time after. Additionally, homeownership is associated with strong financial ties to their current home making them less likely to move than renters. [15]

A Dutch study performed on large-scale longitudinal register data demonstrated adapted mobility behavior in response to neighborhood level “social distance”.The social distance between individuals and neighbors, defined as social differences in education, profession, income, and cultural background was shown to impact the odds of outward mobility from the neighborhood. It follows that individuals with a large social distance with respect to the median social position of the residential neighborhood across various aforementioned dimensions possessed a higher propensity to move than those with a smaller social distance. [16]

Neighborhood ethnic composition was also shown to have an effect on neighborhood out-mobility of individuals. More specifically, it was demonstrated that greater neighborhood proportions of non-western ethnic minorities have a positive impact on the propensity to move, and a further increase was noted for Dutch individuals in neighborhoods with recent increases in the share of non-western ethnic minorities. [17]

Regional economic conditions are also known to possess a modest impact on moving propensity, with individuals less likely to leave “high income regions and metropolitan areas”. [18]

3. Methodology

3.1 Data Sources

Every registered resident in Belgium is given a unique personal identifier, allowing the joining and linkage between different national registers across multiple years. The following national registers were used: The Stock Register containing demographic features including sex, age, civil status, parental IDs, household identifiers, nationality, country of origin, country of birth, household type and position, statistical sector and municipality of residence registered as of January 1st of the reference year; National Death Register consisting of death records covering the entire reference year indicating the date of death ; Educational Register containing records on individual level education status registered as of January 1st of the reference year; Internal Migration Register consisting of individual level migration data over the course of the reference year containing the move date as well as the initial and destination municipality; and Financial Register consisting of individual level income and employment status records registered on January 1st of the reference year.   Demographic, flow(deaths, and internal migration), financial, and educational public registers are linked from the years 2010-2017 and supplemented with statistical sector and municipality data from STATBEL. For the features of interest used to construct the predictive model, there are no missing values.

3.2 Study Sample

150,000 households are initially randomly sampled without replacement in the year 2010, and individuals present across all national registers from the years 2012-2017 are selected. In each year, the entire households of sampled individuals are reselected for the generation of household level features, after which, the initially sampled individuals are retained.

To minimize missingness, only Belgian citizens who remain in Flanders throughout the duration of the time frame were included in the study. Individuals who are at least 18 years old during the first prediction year are included as well, since the moving behavior of children tends to be linked to that of their household guardians. [19] Additionally, only individuals in private households and residing in statistical sectors with a population of greater than 10 are included in the study.
The final study sample contained 198,024 individuals.

3.3 Predictors

The following predictors included in the study were motivated by previous literature on factors influencing moving propensity. For features not included in the aforementioned public registers, they are derived based on existing features.
All time dependent predictors, as well as the target variable, are included for each year in the time horizon 2012-2017.

<u>Target variable</u>: Moving behavior within one year of the reference date.

0: Did not move
1: Moved

3.3.1 Personal Features:The following features are available January 1st of each year.

1. <u>Age_cat:</u> Age is categorized as follows [20]:

    Q0: 0-17
    Q1: 18-26
    Q2: 27-36
    Q3: 37-45
    Q4: 46-53
    Q5: 54-61
    Q6: 54-61
    Q7: 71+

2. <u>Civil Status:</u>

    0: Unmarried
    1: Married
    2: Widow(er)
    3: Divorced

3. <u>Employment Status</u>: Employment status for individuals with net income greater than 400 euros a month:

    0: No paid job
    1: Employed
    2: Employed + other
    3: Self-employed

4. Educational Status: Categorical levels 0-8 indicate educational status at the reference date for each year, with each category corresponding to an increasing level of education.

    0- Early childhood education
    1- Primary Education
    2- Lower Secondary Education
    3- Upper Secondary Education
    4- Post-secondary non-Tertiary Education
    5- Short-cycle tertiary education
    6- Bachelor's degree or equivalent tertiary education level
    7- Masters degree or equivalent tertiary education level
    8- Doctoral degree or equivalent tertiary education level

5. Personal Income: Personal income is computed in line with the modified OECD equivalence scale, whereby a weight is attributed to each person within the household (1.0 to the first adult, 0.5 to the second adult and each subsequent person aged 14 and over, and 0.3 to each child aged under 14). [21] Then adjusted household income is calculated by dividing the total household income by the sum of the equivalence size per household. Quintiles of the adjusted household income are computed for each given year and municipality, and individual incomes are thus classified as "low", "middle" or "high" according to their municipality of residence.

    "Low": Less than the 2nd quintile
    "Middle": Between the 2nd and 3rd quintile
    "High": Greater than the 3rd quintile

6. Origin: Individual origin is derived from an available feature from the public register data, indicating an individual's country of origin with country codes. The following three categories are derived:

    Belgian descent
    Western European descent
    Non-Western European descent

7. Family ties: The presence of family ties for an individual living in a municipality are defined as having at least one parent living within said municipality.

    0- No family ties within municipality
    1- Family ties within municipality

3.3.2 Household Features: The following household features are available or derived from existing features from the reference date of each year.

1. Household Position:

    1- living alone

2- married without children
3- married with children
4- child of/living with a married couple
5- unmarried living together, without children
6- unmarried living together, with children
7- child of/ living with an unmarried couple
8- single parent
9- child of/living with a single parent
10- person not belonging to the family nucleus
11- member of another household type

2. Household Type:

1- living alone
2- married couple without children
3- married couple with children
4- unmarried couple without children
5- unmarried couple with children
6- single parent household
7- other type

3. Number in Household: The total household size is computed at the reference date for each year, and categorized:

1- single person household
2- two person household
3- three person household
4- four person household
5- five or more person household

4. Number of Children in Household: The number of children within each household is computed at the reference date for each year and categorized:

0- no children.
1- one child
2- two children
3-three children
4- four or more children

3.3.3 Statistical Sector Features: The following features are computed for each statistical sector for the reference date of each year. Percentages are then categorized as followed:

1:1-20%
2: 21-40%
3: 41-60%
4: 61:80%
5: 81-100%

1. Percentage of Households with children: The number of households with children including households consisting of a married or unmarried couples with children are counted to generate the total percentage of households with children.
2. Percentage of individuals aged 65 and over:
3. Percentage of individuals by origin**:** Proportion derived from the generated feature *origin*

   Belgian
   Western European
   Non-Western European
4. Percentage of single households.
5. Percentage of single parent households.
6. Percentage of households by income:

   Low income households
   Middle income households
   High income households
7. Median Age Category: The median age is calculated for the Statistical Sector. It is not a percentage.

3.3.4 Municipal Characteristics: The following features are generated for the reference date of each year.

1. Municipality Population Size: Quintiles are computed for population size by municipality, and municipality size is categorized as "low", "middle" and "high"
2. Employment/Unemployment Rate of Municipality**:** Both unemployment and employment rate are categorized into 5 levels corresponding to increasing percentages.

3.3.5 Life History Events: Life history events are generated as dummy variables coded 0 if they did not occur and 1 if they did occur within 1 year of the reference date of each year*.* Life history events are generated for each year included in the study.

1. Change in Civil Status: The civil status at the reference date for reference year and the previous year are used to generate the following:

   0- Divorced
   1- Married
   2- Widowed
2. Change in income status: Generated based on income status at the reference date of each year with respect to the previous year.
3. Death within household: Generated based on whether someone else within the household died within 1 year of the reference date.

4. Birth within household: Generated based on whether a birth in the household took place within 1 year of the reference date.
5. Change in employment status: The employment status at the reference date at the year and the previous year are used to generate the following features:

   0- Unemployed

   1- Employed
6. Change in Household position: Derived from the household position at the reference date of each year with respect to the household position at the reference date of the previous year
7. Change in Household Type: Derived from the household type at the reference date of each year with respect to the household type at the reference date of the previous year
8. Someone within the household moved: Generated based on whether someone else within the household moved houses within 1 year of the reference date.
9. Cumulative sum up until each year is included for moves, moves within the household, births, and deaths.

3.3.5 Time since life history events: The time since life history events are computed with respect to the reference date for each year starting from 2010.

1. Predictors for the prediction year *t=2016, 2017* contain the following categories, excluding events that take place after the reference date of the prediction year *t*.

   0-6 months since event date

   6-12 months since event date

   12-18 months since event date

   18-24 months since event date

   2 + years since event date

   No change: Event didn't take place up until the reference date
2. Predictors for years *t-1…t-4*, which include the time since and after the reference date of each year and are categorized as follows.

   0-3 months since reference date

   3-6 months since reference date

   6-12 months since reference date

   0-6 months since event date

   6-12 months since event date

   12-18 months since event date

   18-24 months since event date

   2 + years since event date

   No change: Event didn't take place up until the reference date

3. Time between the last move and last death within household, birth within household, and someone else within the household moved are derived for the years *t-1…t-4* and are categorized as follows:

   0-3 months since since move
   3-6 months since since move
   6-12 months since since move
   12-18 months since move
   18-24 months since move
   2 + years since move
   0-6 months since event date
   6-12 months since event date
   12-18 months since event date
   18-24 months since event date
   2 + years since event date
   No change: Event didn't take place up until the reference date

3.4 Sampling Methods to Rectify Class Imbalance

Table 1: *Class Distribution Target Variable*

	Stayers	Movers	Movers (%)
Prediction year t	189545	8479	4.28
Prediction year t+1	188458	9566	4.83

Class imbalance in binary classification problems, particularly when the minority class is the class of interest, can be problematic for predictive modeling since many machine learning algorithms assume equal class distributions. As the learning process of most classification algorithms is often biased towards the prevalent class, instances belonging to the minority class are more often misclassified resulting in poor predictive performance. Given the skewed class distribution of the feature of interest, it is indispensable to address the class imbalance prior to modeling. [22]

There are numerous sampling methods aimed at rectifying skewed class distributions of the target variable. One such method of balancing data is random under sampling without replacement, whereby samples from the majority class are removed at random to match the size of the minority class. A significant drawback of under sampling is loss of information, since it may remove potentially useful data essential for the learning process, resulting in poorer predictive performance. [23] In the case of the present study, where the ratio of the minority class to the majority class is less than 1:20, random undersampling of the training set would result in a significant loss of information for the learning process.

Another is random oversampling with replacement, whereby the minority class is oversampled to match the size of the majority class. Since oversampling effectively makes duplicates of existing minority samples to balance data, the chances of overfitting to the training data become more likely. Additionally, oversampling results in larger data samples, ultimately increasing computational cost. [24]

Generation of new artificial data for the minority class is another method aimed at rectifying the loss of information of random undersampling and the increased likelihood of overfitting of random oversampling with replacement.

Random OverSampling Examples (ROSE), generates new synthetic instances of the minority class from an estimate of the conditional density underlying the data based on a smoothed bootstrap form of re-sampling the data. It has been shown to outperform predictions based on unbalanced, as well as random oversampled and random undersampled training data. [25]

To inform the decision on the sampling method of choice to rectify the class imbalance, a preliminary classification decision tree to predict moving behavior in year t is performed on the training data treated with random undersampling without replacement, random oversampling with replacement, and the ROSE method and evaluated on the validation data. The sampling method with the best performance is then chosen for model estimation, the results of which are seen in Figure 3.

To minimize the computational effort of the training process, the oversampled and ROSE training data are limited to 50,000 observations, entailing an additional downsampling without replacement of the majority class. Undersampling and oversampling of the training data was performed using the “ROSE” package in R. [26]

3.5 Evaluation Metrics:

All quality measures assess the predictive performance of each model based on predicted vs. actual moving behavior, and are based on the *confusion matrix,* whereby the predicted probability that an individual moves is translated into the binary target variable “stayed” or “moved”. A classification threshold value is used, whereby probabilities above the threshold are classified as “moved”, and probabilities below the threshold are classified as “stayed”.

Table 2: *Confusion Matrix*

Predicted Observed	Stayed	Moved
Stayed	TN	FP
Moved	FN	TP

TP: true positives, the number of correctly classified movers.
TN: true negatives, the number of correctly classified stayers.
FN: false negatives, the number of movers misclassified as stayers.
FP: false positives, the number of stayers misclassified as movers.

$$Accuracy = \frac{TN+TP}{TN+FN+TP+FP}$$

A common evaluation metric is accuracy, which is defined as the fraction of the total predictions that are correct. While commonly used, accuracy is sensitive to class imbalance, and therefore misleading as a standalone metric for assessing model performance.

$$TPR = \frac{TP}{TP+FN} \qquad TNR = \frac{TN}{TN+FP} \qquad FPR = 1 - TNR$$

The “sensitivity”, or True Positive Rate (TPR), indicates the proportion of movers that are correctly predicted by the model. The “specificity”, or True Negative Rate (TNR), is the proportion of stayers that are correctly predicted, while the False Positive Rate (FPR), indicates the proportion of stayers misclassified as movers. Given the imbalanced class distribution of movers and stayers, it follows that TPR, TNR, and FPR are more informative regarding the predictive performance of the derived models.

The receiver operating characteristic (ROC) curve, is a visualization of the trade off between the TPR and FPR at different classification thresholds, and provides an overall representation of the predictive performance of the classification model. The area under the ROC Curve, or “AUC”, provides an aggregate measure of performance across all possible classification thresholds and indicates the ability of the model to distinguish between movers and stayers.

$$Balanced\ Accuracy\ =\ \frac{TPR+TNR}{2}$$

Finally, *balanced accuracy*, which takes into account class imbalance, and is defined as the average between the TPR and TNR. It follows that given the unequal class distributions of movers and stayers, this evaluation metric is appropriate.

3.6 Machine Learning Algorithms

The 3 model algorithms, elastic net penalized logistic regression , random forest, and XGBoost are widely studied and commonly used machine learning models. They are chosen due to their performance with high dimensional data and multi-collinearity amongst predictors.

3.6.1 Elastic Net Penalized Logistic Regression

The performance of classical logistic regression in binary classification problems tends to suffer in the face of high dimensionality due to increased risk of over-fitting.

Penalized logistic regression entails an imposed penalty to logistic regression for such high dimensionality, and results in shrinking coefficients of less contributive predictors towards zero, producing a more stable classifier. This is made possible by two types of regularization: l_2 and l_1.

The l_2 ridge penalty is defined as: $\lambda \sum_{j=1}^{p} \hat{\beta}_j^2$

The l_1 LASSO (Least Absolute Shrinkage and Selection Operator) penalty is defined as: $\lambda \sum_{j=1}^{p} \left|\hat{\beta}_j\right|$

With λ the penalty parameter, emphasizing how much weight is given to the penalty.

l_1 regularization tends to shrink less contributive features to 0, resulting in a more parsimonious model, making it ideal dimensionality reduction when dealing with high dimensional data. With l_2 regularization, on the other hand, the shrinkage of less contributive features does not fully achieve 0, reducing the effect of highly correlated predictors without removing them from the model. The elastic net penalized logistic regression (elastic net), combines the l_1 and $l_2$penalties of the LASSO and Ridge methods via the new α parameter, with α=0 giving way to the l_2 regularization of the Ridge regression and α=1 resulting in the l_1 regularization or LASSO regression. The estimates are produced from the minimization of the following loss function:

$$L_{enet} = L_{log} + \lambda \sum_{j=1}^{p} (\alpha \hat{\beta}_j^2 + (1 - \alpha) \left| \hat{\beta}_j \right|)$$

With "Logistic Loss" function, L_{log}, and $0 \leq \alpha \leq 1$.

Both the α and λ are tuned on the training data, with optimal values chosen for the final model. The ranking of features by importance is obtained through absolute magnitude of the predictor coefficient produced by the elastic net. [27]

3.6.2 Random Forest

Random Forest is an ensemble-based learning algorithm made up of an ensemble of decision trees, incorporating bootstrap sampling of observations and random subselection of predictors at each decision node. In the case of classification, the Random Forest algorithm employs numerous de-correlated trees to compute majority votes in the terminal leaf nodes for the outcome class. The Random Forest algorithm has several advantages, including handling high dimensional and large dataframes composed of quantitative and qualitative predictors. Another advantage is its ability to navigate high correlation amongst predictors which may become problematic for non tree-like learning algorithms. [28]

Random Forest Algorithm:

```
1. For $b = 1$ to $B$:
    a. Draw a bootstrap sample Z* of size N from the training data
    b. Grow a random forest tree $T_b$ to the bootstrapped data, by recursively
       repeating the following steps for each terminal node of the tree, until the
       minimum node size $n_{min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables
        ii. Pick the best variable/split-point among the $m$
        iii. Split the node into two daughter nodes
2. Output the ensemble of trees $\{Tb\}_1^B$
   To make a prediction at a new point $x$:
```

Let $\widehat{C}_b(x)$ be the class prediction of the b^{th} random-forest tree. Then:

$$\widehat{C}^{B}_{rf}(x) = majority\ vote \left\{\widehat{C}_b(x)\right\}^{B}_{1}$$ [29]

The following random forest hyperparameters are tuned:

1. *min_n*: the minimum number of observations in a node required for the node to be split.
2. *mtry:* the number of predictors randomly sampled at each split.

Both are important in the prevention of overfitting, while the number of trees is held constant at the default value of 500.

Feature importance for the random forest model is based on the mean decrease in the Gini coefficient or Gini index, which indicates the extent to which each feature contributes to the homogeneity of the nodes and leaves in the resulting random forest.

$$Gini\ Index = \sum_{k=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk})$$

With $\widehat{p}_{mk}$ the proportion of training observations at node m in class k, and K the total number of classes. [30] Impurity importance for each feature is computed by the sum of all impurity decrease measures of all nodes in the forest at which a split on said feature was conducted, normalized by the number of trees.

3.6.3 XGBoost

The XGBoost (Extreme Gradient Boosting) algorithm, another ensemble based learning algorithm, expands on Gradient Tree Boosting, a tree boosting technique to reduce the loss function using additive learning of the weak learners, in this case decision trees. Unlike the random forests which build individual decision trees independently, Gradient Boosting combines weak learners sequentially so that each new tree corrects the errors of the previous one.

The objective function is defined as:

$$O(\Theta) = \sum_{i=1}^{N} L(y_i, \widehat{y}_i^{(t)}) + \sum_{k=1}^{K} \lambda(f_k)$$

With L the loss function $L(y_i, \widehat{y}_i^{(t)}) = -[y_i log(\widehat{y}_i^{(t)}) + (1 - y_i)log(1 - \widehat{y}_i^{(t)})]$, λ the regularization parameter, and $\lambda(f_k)$ is the complexity of the tree f_k. y_i the class label at instance i, $\hat{y}_i^{(t)}$ the prediction of i^{th} instance at iteration t. K corresponds to the number of trees and N the number of trees.

First and second order optimization to find local minima are accomplished through the gradient g_i and hessian h_i, which integrates $\hat{y}_i^{(t-1)}$, the prediction of i^{th} instance at iteration $t - 1$.

$$g_i = \partial_{\hat{y}_i^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)})$$

The final objective function becomes at iteration t:

$$O(\Theta) = \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \lambda(f_t)$$ [31,32]

The following XGBoost hyperparameters are tuned:

1. The *Learn Rate*: the step size shrinking used to descend the gradient
2. *Loss Reduction*: the minimum loss reduction required to make a split
3. *Min_n,* the minimum sample per node require for split,
4. *M_try,* the number of predictors sampled for each tree,
5. *Sample size:* the proportion of observations randomly sampled for each tree.
6. *Tree depth:* the maximum depth for each tree.

The top 30 ranked predictors for the XGBoost model is based on relative importance or "relevance" of predictors. The relevance of X_i in separating the class k observations from other classes is defined as:

$$I_{lk}^2 = \frac{1}{M} \sum_{m=1}^{M} I_{lk}^2 \left(T_{km} \right)$$

The overall relevance of X_i is obtained by averaging over all classes.

$$I_l^2 = \frac{1}{K} \sum_{k=1}^{K} I_{lk}^2$$ [33]

3.7 Prediction Methods

Table 3: *Train, Validation and Test Split*

		Sample Size	2012	2013	2014	2015	2016	2017
90%	TRAIN	178,221 (50,000 after ROSE)	t-4	t-3	t-2	t-1	t	
10%	VALIDATE	19,803	t-4	t-3	t-2	t-1	t	
	TEST	19,803		t-3	t-2	t-1	t	t+1

The study sample of 198,024 individuals was split using stratified random sampling into a 90% training set and a 10% validation set for the years t-4…t for prediction year t=2016 to assure equal class distributions of the target variable. The test set, reserved for the final model validation, consists of the same individuals present in the validation set but for the years t-4…t for prediction year t=2017.

All features known up to a reference date of year *t* were used to predict moving behavior within 1 year of the reference date. January 1st, 2016 was used as a reference date to optimize and train the models, with further optimization performed using the validation set. Trained models were applied to the test set to assess how well they generalize to another year.

For the random forest model derivation, there were 326 total predictors. For the XGBoost and eNET models, all categorical predictors were converted to dummy variables through one hot encoding and the sparse matrix of predictors were generated, resulting in 1,414 total predictors.

For each prediction method, the optimal set of hyperparameters are grid searched via 10-fold cross validation on the training set treated with the ROSE upsampling method. The initial hyperparameter space for each model is chosen using the train AUC as the evaluation metric.

The initial grid search for the XGBoost and random forest models are performed using the R package “TidyModels” [34], while the elastic net, random forest, and XGBoost models were performed using the “glmnet” [35], “randomForest” [36], and “xgboost” [37] packages.

Additional hyperparameter tuning for select hyperparameters is performed for the random forest and XGBoost models and evaluated on the validation set to avoid overfitting, after which the final hyperparameter space is selected.

Then cutoff probability thresholds for classifying an individual as a “mover” or “stayer” are chosen for the final predictive models for each machine learning algorithm based on a trade off between the TPR and the FPR. Probability thresholds yielding a minimal acceptable range of TPR values of 60-80% are considered. Within this range, the median threshold value c^{*}_{med} as well as the threshold resulting in the minimal FPR $c^{*}_{min\,FPR}$ are extracted from the model predictions on the validation set and then applied to model fit on the validation and test dataset. The performance of each model is evaluated through the accuracy, TPR, TNR, FPR, and balanced accuracy.

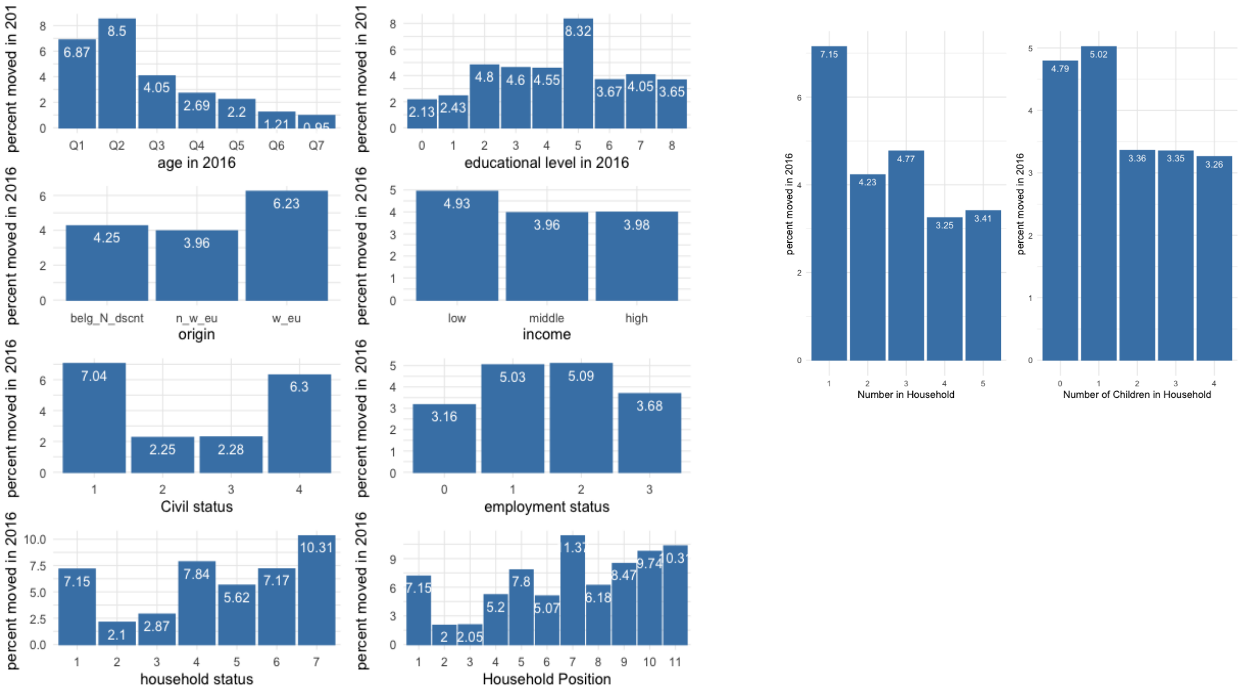
Top 30 predictors were reported separately for the elastic net, random forest, and XGBoost model. The selection of the top predictors were based on the absolute magnitude of predictor coefficient for the elastic net, and predictor importance score for random forest and gradient boosting models.

Finally, to gain a better understanding of prediction errors, two classification decision trees are built based on the class predictions. The first, to classify observations that are correctly or incorrectly predicted. The second, classifying false positives and false negatives among misclassified observations. Given the high dimensionality of the data prepared for the eNET logistic regression and XGBoost models, the classification decision trees are performed solely for the final random forest model.

4. Results

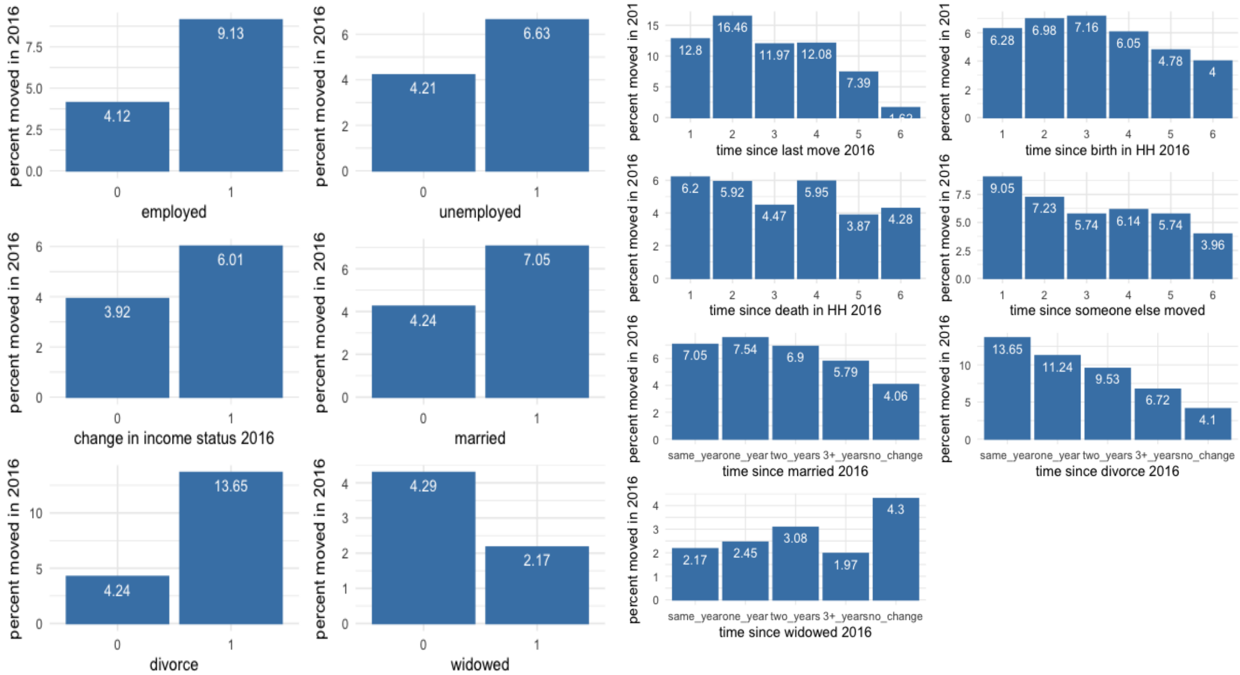
4.1 Exploratory Data Analysis

Figure 1: Relative Frequencies of movers in year t=2016, Personal and Household Characteristics



In Figure 1, we see the relative frequencies of individuals who moved in year t=2016 and select personal and household characteristics. Consistent with the literature, individuals aged 18-26 (Q1), and 27-36 (Q2) are the most mobile, with relative frequencies of 6.87% and 8.5% respectively. Individuals possessing a short-cycle tertiary education have the highest mobility in year t (8.32%), with increasing education levels corresponding to less mobility. Individuals of western european descent have higher relative moving frequency as compared to individuals of Belgian descent or non-western european descent. Low income individuals, it follows, possess marginally higher mobility as compared to middle and high income individuals, while unmarried and divorced individuals are more mobile than the married or widowed in year t. Those who were employed or employed with another supplementary source of income had slightly higher mobility compared to individuals who were unemployed or self-employed in year t. Individuals who belong to a single person household, a household composed of an unmarried couple without children, or other household type, are the most mobile as compared to other household compositions. We also see that the proportion of movers decreases with household size and number of children.

Figure 2: Relative Frequencies of movers in year t=2016, life history events and time since life history event



In Figure 2, we see the relative frequencies of individuals who moved in the prediction year t, and select life history events, as well as the time since the last life history event for the same year. For individuals who found employment within the previous year, the proportion of movers is more than double in year t. For those who became unemployed in the previous year, the relative frequency is higher than those who did not become unemployed. Individuals who experienced a change of income status within the previous year appear to be more mobile in year t. Getting married within the previous year results in a higher mobility in year t, and union dissolution results in nearly triple the proportion of movers in year t as compared to those who did not experience a divorce. Becoming widowed demonstrates the opposite effect, with widowers being half as mobile as individuals who were not widowed in the previous year.

When looking at the time since the last move, the highest relative frequency of movers occurs when the last move occurred within 6-12 months of the reference date for year *t* (16.46%), followed by 0-6 months of the reference date (12.8%). For individuals whose last move was more than 2 years ago, the relative percentage of movers drops to 7.39%, and for those who had not moved since 2010, the relative percentage is 1.62%. This indicates that a higher proportion of individuals tend to move in year t when a previous move occurred closer to the reference date, and moreover, individuals who had not moved since 2010 appear less mobile in year *t.*

A similar pattern is observed for the time since someone else moved, with the highest proportion of movers occuring when someone else within the household moved within 0-6 months of the reference date for year *t* (9.05%). The relative frequency decreases with increasing distance in time from the reference date since someone else within the household moved, with the lowest for individuals who did not have someone leave their household in the last 5 years. This suggests higher mobility for individuals experiencing a recent change in household composition, and since a change in household composition can often be an unanticipated event this is consistent with the literature.

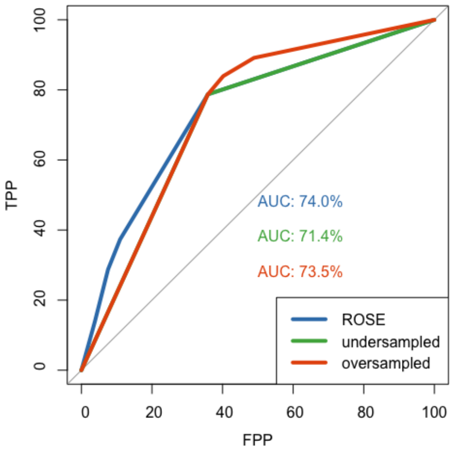
The highest proportion of movers occurs when the last birth within the household was between 12-18 months before the reference date of year *t* (7.16%), and lowest when no birth occurred in the last five years (4%). Though the difference in the proportion of movers is not substantial between time spans 0-6 months, 6-12 months, and 12-18 months before the reference date, there is a drop in relative frequency as time since the last birth within the household increases with respect to the reference date after 12-18 months.

Time since the last death within the household does not appear to result in any observable pattern with respect to time since event and mobility in year *t.* Nonetheless, the highest proportion of movers occurs when the death within the household took place 0-6 months before the reference date, though the proportion is not substantially greater than individuals with no death within the household in the last 5 years.

The time since divorce possesses highest proportion of movers for union dissolutions taking place within the same year (13.65%) and steadily decreases with every additional year since the event with respect to the reference date of year *t,* and appears consistent with the literature demonstrating that a union dissolution, an often unanticipated event, will trigger a move. The time since union formation, or marriage, does not show the same pattern in moving proportions, likely because union formations are anticipated events, and likely preclude a change in housing.

4.2 Classification Decision Tree and Sampling Method for Training Data

Figure 3: Classification Decision Tree Training Set Sampling Method

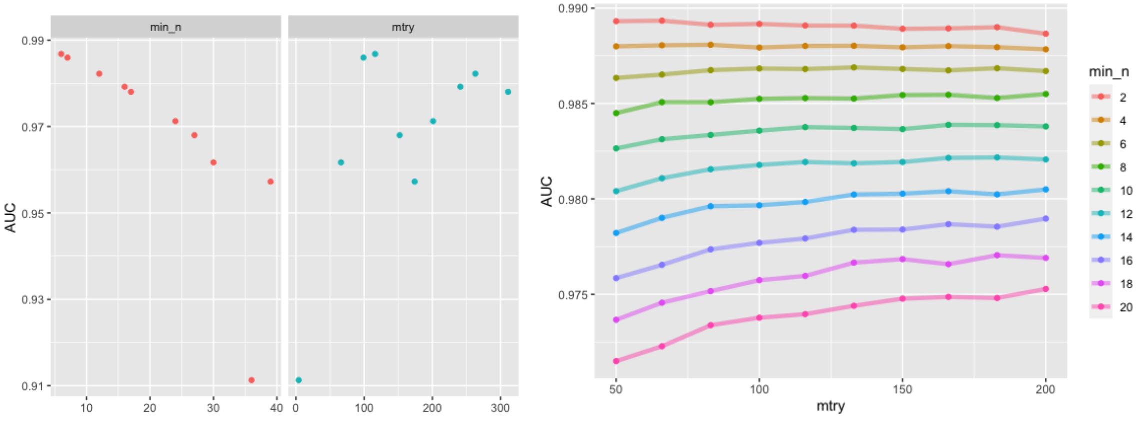


Sampling Method	AUC with 95% CI
Undersampled	71.40 (70.04-72.85)
Oversampled	73.5 (72.24-74.77)
ROSE	74.00 (72.37-75.55)

The preliminary Classification Decision Trees classifying movers in year t with undersampled without replacement, oversampled with replacement, and ROSE training data and their evaluation on the validation set are displayed in Figure 3. The mean AUC for the decision tree model trained on the ROSE method is the highest as compared to the other two sampling methods. Though the difference between mean AUCs for each method is not statistically significant, nonetheless, the results indicate a modest improvement in terms of predictive performance with the ROSE method. The literature on the superior performance of ROSE with respect to undersampled and oversampled methods, provides additional motivation for favoring the ROSE sampling method. As a result, the random forest, XGBoost, and penalized logistic regression models are built and trained on the ROSE training data.

4.3 Hyperparameter Tuning and Model Building

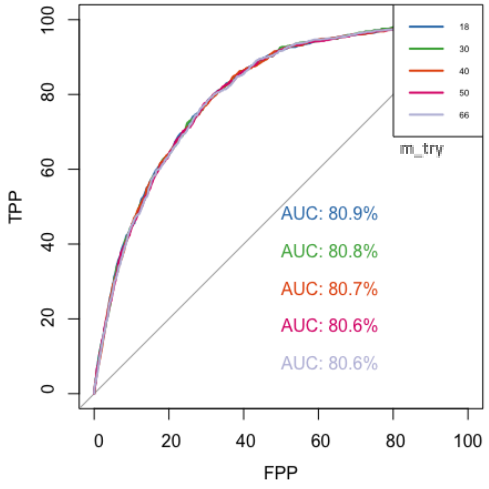
Figure 4: Hyperparameter Grid Search Random Forest.



The hyperparameter grid search for the random forest model is performed on the training data to find optimal values for the random forest hyperparameters. The results of the second grid search on a target range of *min_n* and *mtry* values indicate an optimal AUC value of ≈0.99 for *min_n* of 2 and *mtry* of 66. Overall, the model produces near perfect fit on the training data, and therefore, near perfect discrimination between stayers and movers.

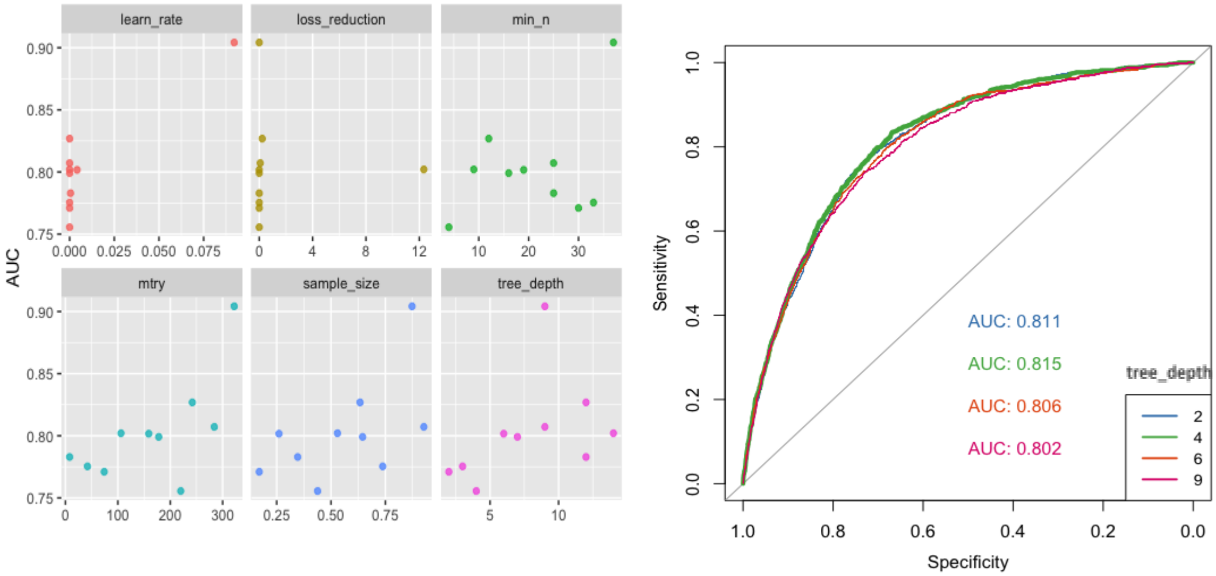
To avoid overfitting, the number of variables sampled at each split is further tuned, while keeping all other parameters constant, and evaluated on the validation set, the results of which are seen in Figure 5.

Figure 5: Hyperparameter Tuning Random Forest *mtry.*



The random forest model is retrained with all other hyperparameters held constant and decreasing *mtry* values, effectively decreasing tree complexity and mitigating the risk of overfitting. The AUC for trained models with *mtry* of 18, 30, 40, 55, and 66, the original optimal value identified in the hyperparameter grid search range from 0.809-0.806 respectively after being evaluated on the validation set. Already, the results indicate a substantial amount of overfitting, with a drop in AUC of around 18% as compared to the initial fit on the training data. Given the unsubstantial difference in the AUC score between models trained with varying *mtry* values, the smallest value for *mtry* is chosen for the final model, which corresponds to random forest model with the least complexity, and therefore decreases the likelihood of overfitting when evaluated on the test set.

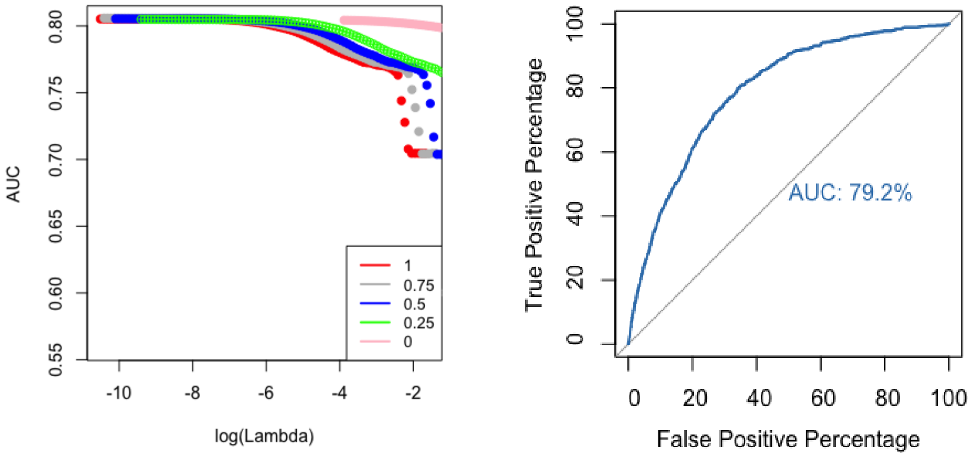
Figure 6: Hyperparameter Grid Search and Tuning XGBoost.



The optimal values identified from the hyperparameter grid search for the XGBoost model produce an AUC of ≈ 0.91, indicating strong discriminatory accuracy between stayers and movers. The optimal hyperparameters are used to retrain the model, but now with decreasing values of *tree_depth*, starting from the optimal value identified from the grid search of 9, and evaluated on the validation set. As with the random forest model, there is evidence of overfitting, given the drop in validation AUC of 9-10%.

The validation AUCs for decreasing values of *tree_depth*, and therefore decreasing model complexity, are not substantially different. Nonetheless, the XGBoost model trained with a *tree_depth* of 4 produces highest mean validation AUC of 0.815 and a slightly steeper ROC curve, and since this corresponds to a lesser degree of model complexity as compared to the initial *tree_depth* of 9 as identified in the grid search, it is selected for the final model.

Figure 7: Hyperparameter Tuning eNET.



The hyperparameter tuning results for eNET model produce similar results on the training dataset, with varying values for the alpha parameter (0, 0.25, 0.5, 0.75 and 1.0) yielding a mean AUC of 0.8043, 0.8050, 0.8054, 0.8058, and 0.8052 respectively at their respective optimal lambda values. The model with α=0.75 and minimum λ = 0.0000301 is chosen, and therefore represents a LASSO-Ridge mix with a greater weight given to the LASSO regression, whereby coefficients are to a greater extent pushed to zero resulting in a more parsimonious model.

The retrained model with the aforementioned hyperparameter values is evaluated on the validation set, the results of which are seen on the right hand plot in Figure 7. The fit on the validation set does not result in a large drop in AUC, decreasing from 0.8058 to 0.792, and indicates little evidence of overfitting on the validation data.

4.4 Prediction Results

Table 4: *Model performance metrics at median and minimum FPR risk thresholds for predicting moving behavior in year t.*

Model	AUC (%) with 95% CI	ACC (%) with 95% CI	TPR (%)	TNR (%)	FPR (%)	bACC(%)
eNET c^{*}_{med}	79.2 (77.75-80.57)	73.02 (72.40, 73.64)	72.14	73.06	26.94	72.60
eNET $c^{*}_{min\,FPR}$		79.51 (78.94, 80.07)	60.12	80.39	19.61	70.26
RF c^{*}_{med}	80.9 (0.7949-0.8227)	75.58 (74.97, 76.18)	70.29	75.82	24.18	73.06
RF $c^{*}_{min\,FPR}$		82.08 (81.54, 82.61)	60.12	83.08	16.92	71.59
XGBoost c^{*}_{med}	81.5(80.11-82.80)	76.55 (75.96, 77.14)	71.56	76.78	23.22	74.17
XGBoost $c^{*}_{min\,FPR}$		82.90 (82.37, 83.42)	60.12	83.94	16.06	72.03

Table 4 shows the prediction quality of each final model on the validation dataset evaluated at the median and minimum FPR thresholds. Overall, the prediction quality does not vary significantly between models, with small differences seen with differing cutoff thresholds. The mean validation AUC is highest with the XGBoost model (81.5%), though not substantially higher than the other two methods. This indicates nonetheless, that the XGBoost model is slightly better at distinguishing between movers and stayers. Generally, models evaluated at $c^{*}_{min\,FPR}$ result in a higher overall accuracy as compared to models evaluated at c^{*}_{med}, which is to be expected as the threshold reflects when the FPR is minimal, and by extension, the TNR is maximized conditional on the TPR being between 60 and 80%. It follows that given the class imbalance in the data, minimizing the FPR will result in an overall higher accuracy, with more stayers, the majority class, being correctly predicted. While all three models result in the same proportion of movers that were correctly classified (60.12%), the XGBoost model boasts the highest proportion of correctly classified stayers (83.94%), and by extension, the minimal percentage of incorrectly classified stayers (16.06%). It follows that the XGBoost model with the threshold $c^{*}_{min\,FPR}$ has the highest mean balanced accuracy of 72.03%, which indicates good predictive performance of moving behavior in year t, though not strongly outperforming the eNET and random forest models. Similar results are seen for models evaluated at c^*_{med}, with the XGBoost model resulting in the highest mean accuracy of 76.55%. The eNET model however, results in a higher TPR, and therefore slightly outperforms the random forest and XGBoost models in terms of classifying movers (72.14%), while the XGBoost models performs marginally better in terms of classifying stayers (76.78%), and maintains the highest mean balanced accuracy (74.17%), and therefore better predictive performance.

Table 5: *Model performance metrics at median and minimum FPR risk thresholds for predicting moving behavior in year t+1.*

Model	AUC (%) with 95% CI	ACC (%) with 95% CI	TPR (%)	TNR (%)	FPR (%)	bACC(%)
eNET c^*_{med}	77.3% (75.93-78.69)	69.75 (69.10, 70.39)	71.34	69.66	30.34	70.50
eNET $c^*_{min\ FPR}$		76.37 (75.77, 76.96)	60.47	77.23	22.77	68.85
RF c^*_{med}	76.69%. (0.7529-0.7809)	73.45 (72.83, 74.06)	64.90	73.91	26.09	69.40
RF $c^*_{min\ FPR}$		74.40 (73.78, 75.00)	62.15	75.06	24.94	68.61
XGBoost c^*_{med}	77.17%(75.76-78.58)	74.92 (74.31, 75.52)	62.94	75.56	24.44	69.25
XGBoost $c^*_{min\ FPR}$		81.76 (81.21, 82.29)	52.17	83.35	16.65	67.76

Table 5 shows the prediction quality for each model generalized to one additional year, evaluated at c^*_{med} and $c^*_{min\ FPR}$ from the model predictions on the validation set. Across the three models and their corresponding risk thresholds, there is an overall moderate drop in mean accuracy, indicating that models trained on t-u for u=0,...,4 do not generalize perfectly to the test data consisting of years t-u for u=-1,..,3. This provides evidence that the predictive models possess moderate variance and overfit to the years t-u for u=0,...,4. Nonetheless, when minimizing the FPR, the XGBoost outperforms the random forest and eNET models in terms of predicting the highest number of observations correctly (81.76%). It follows that 83.35% of stayers are correctly predicted, though this comes at the expense of the TPR, however, with only 52% of movers correctly predicted. At c^*_{med}, the eNET model possesses the highest TPR, correctly predicting 71.34% of movers, though this comes at the cost of a higher rate of stayers being misclassified as movers (30.34%). It follows that the eNET model with c^{*}_{med} also achieves the highest mean balanced accuracy (70.5%), though the metric does not vary significantly between methods and risk thresholds.

Despite the moderate variance, prediction quality of moving behavior in year t+1 is within acceptable range for good predictive performance across the elastic net, random forest model, and XGBoost model.

4.5 Feature Importance

Figure 8: Feature Importance, random forest

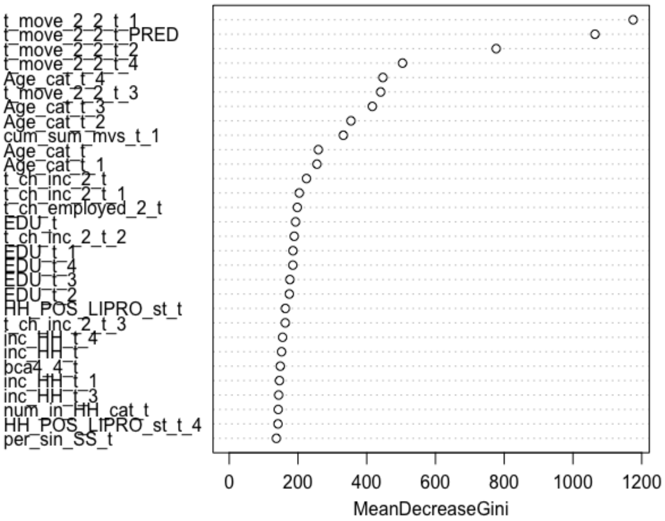
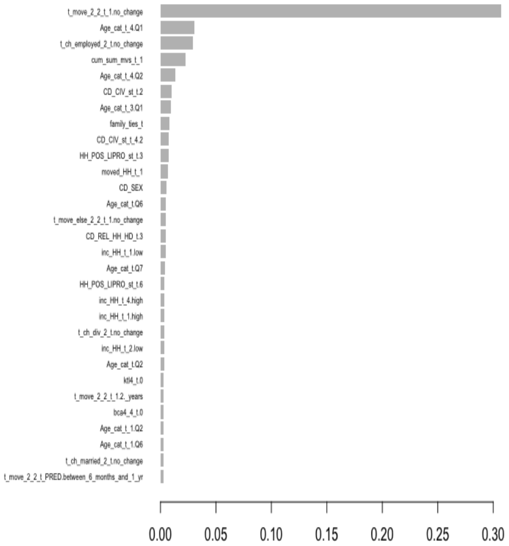


Figure 8 shows the top 30 predictors with the highest importance based on the final elastic net, random forest model, and XGBoost model. The ranking of features was dependent on the method, with certain features overlapping between methods, while others selected uniquely in one model.

According to the random forest model, the time since the last move for each year made up 5 of the top 6 predictors in terms of feature importance, with time since last move at year t-1, year t, and year t-2 resulting a substantially greater mean decrease in the Gini coefficient as compared to the other predictors. Other important features include age for years t-u for u=0,...,4, as well as time since last change in personal income status for years t-u for u=0,..,2, education status for years t-u for u=0,...,4, time since becoming employed for year t, household position at years t and t-4, employment status at year t, personal income status for year t-u for u=0,1,3,4, number within the household, the neighborhood percentage of single households at year t, and the cumulative sum of moves at year t-1. For the random forest model, the top 30 ranked predictors, (making up nearly 9% of the total number of features) is largely made up of individual and household level characteristics. Statistical sector and municipal level characteristics, do however make the top 100 ranked predictors (30% of the total features), with the municipality size and the unemployment/employment rates for each year achieving a mean Gini coefficient of greater than 100, indicating that their relative contribution is not negligible.

Figure 9: Feature importance, XGBoost



The top predictor for the XGBoost model is the time since the last move with no change in the last 5 years at year t-1, followed by a large drop in mean gain for the remaining features whose mean gain is less than 0.05. Nonetheless, the age range of 18-26 and 27-36 at year t-4, the cumulative sum of moves in year t-1, and the time since becoming employed with no change in the last 5 years in year t possess relatively high predictive importance for classifying movers in the XGBoost model.

The remaining top features consist mostly of personal, household characteristics, as well as time since life history events. The presence of family ties is the 8th most important feature for predicting moving behavior, which did not rank high with the random forest model. Given the low mean gain of the top predictors following the time since last move with no change in the last 5 years at year t-1, as well as the large number of features, it is difficult to extrapolate any meaningful causal relationships with moving behavior.

Figure 10: Feature Importance elastic net

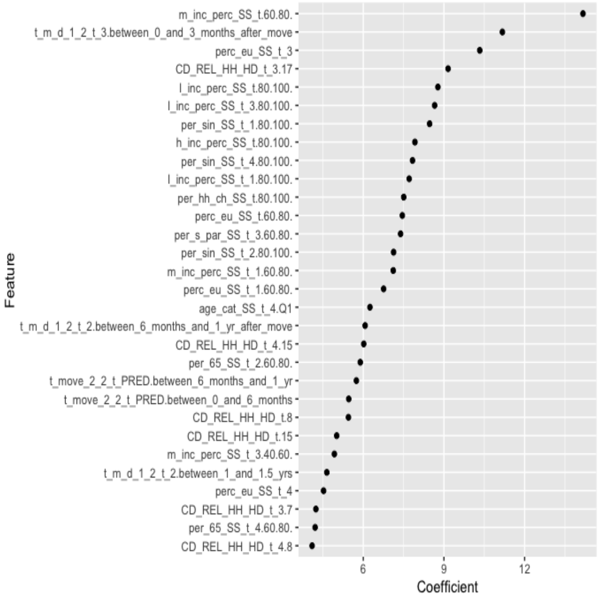
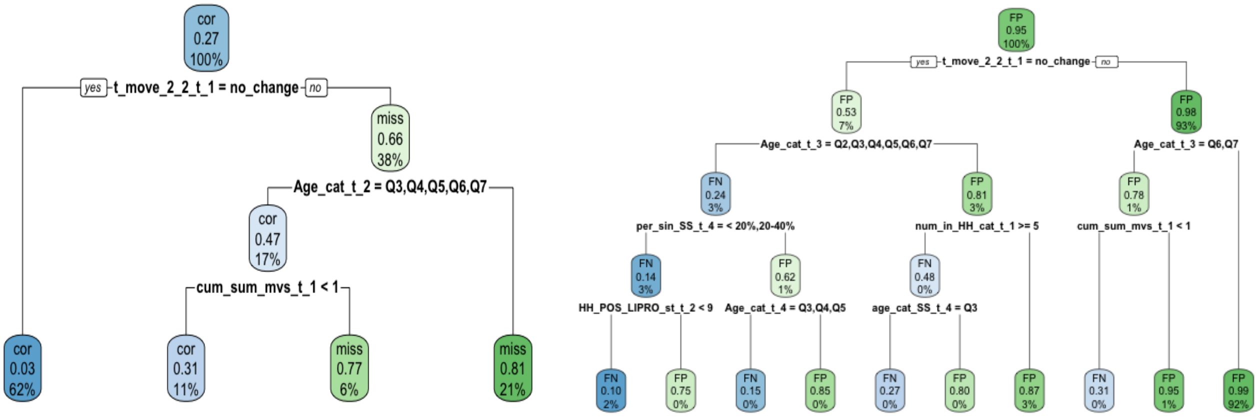


Figure 10 shows the top 30 predictors identified by the elastic net logistic regression, ranked according to the absolute magnitude of the coefficient. Unlike the random forest and XGBoost model, statistical sector level predictors feature prominently, with 20 of the 30 top predictors related to the statistical sector. Additional features with high predictive performance unique to the elastic net logistic regression include the time between a death within the household and the last move at years t-u for u=2,3, the household position relative to the head of the household at years t-u for u=0,2,3,4. The time since the last move in prediction year t, also possesses high relative predictive performance in the elastic net logistic regression.

Given the method dependent results, the ranking of the features by importance do not provide a clear interpretation of causal relationships between predictors and moving behavior. Nonetheless, the data and the features included appear sufficient for classification in year t as well as prediction in year t+1.

4.6 Misclassification and False Positives

Figure 11: Decision Tree Misclassification and False Positives



In the left plot of Figure 11 is the classification decision tree for observations correctly or incorrectly classified by the final random forest model evaluated on the validation set at c_{med}^{*}. Those with no move within the last 5 years at year t-1 were correctly classified by the random forest model with a 3% probability of being misclassified, representing 62% of total individuals. Those with a move within the last 5 years at year t-1 as well as belonging to the age category Q1 or Q2, had a 81% probability of being misclassified, representing 21% of the total number of observations. On the other hand, individuals with a move within the last 5 years at year t-1,belonging to the age categories Q3:Q7, and a cumulative sum of moves in year t-1 greater than 1 possessed a 77% probability of misclassification.

The right plot of Figure 11 shows the classification decision tree for False Positives or False Negatives among misclassified observations by the final random forest model evaluated on the validation set at c_{med}^{*}. Individuals with a move within the last 5 years had a 98% probability of being misclassified as movers when they were in fact stayers, and if they belonged to an age category at year t-3 of Q1:Q5, the probability increases to 99%, accounting for 92% of total misclassified individuals. Those with the lowest probability of being misclassified as movers by the model, were individuals without a move within the last 5 years, belonging to age categories Q2:Q7 at year t-3, living in a statistical sector with less than 20% or between 20-40% of single households, and having a household position other than “child of/living with a single parent”.

According to the classification decision tree results, the predictor time since last move at year t-1 appears to be important in determining the prediction quality of the random forest model, which has the highest probability of correctly classifying individuals as movers or stayers when there has not been a move since 2010. Furthermore, the highest probability that the random forest model misclassifies stayers as movers, occurs when the time since last move at year t-1 is anything other than no move in the last 5 years. The age category at year t-2 or t-3, also appear to be confounding for the random forest in predicting moving behavior. Nonetheless, the results invite reflection on the overall structure of the features present in the above decision trees, and present an opportunity to interrogate the use of continuous predictors in lieu of categorical ones (in the case of time since last event etc.)

5. Discussion

5.1 Summary of Findings

The research question of this study was whether or not moving behavior within one year of the reference date can be predicted using public register data, drawing upon personal, household, neighborhood, and municipal level characteristics as well as life history events for years t=2016,2017. The greater implications of this study, include the viability of secondary data, in this case Belgian public register data, as an alternative to survey questionnaires in predicting moving behavior. This project aimed to expand upon the Dutch study on predicting moving propensity using public register data.

Following hyperparameter tuning via 10-fold cross validation and evaluation on the out of sample validation set, the final elastic net, random forest, and XGBoost models achieved good discrimination between stayers and movers, achieving mean AUCs of 79.2-81.5% respectively. The models generalize reasonably well to one year into the future, with test AUCs for the elastic net, random forest, and XGBoost model of 77.30%, 76.69%, and 77.17% respectively. There is indeed a drop in predictive performance across the final three models when generalized to year t+1. This is likely due to overfitting of the model to the years t-u for u=0,..,4 and t=2016 of the training data, though the extent of the overfitting appears moderate. This could be a result of the effect of social and economic factors not included in the study that impact moving behavior, such as economic crises or fluctuations in the housing market.

The ranking of features by importance differed between models, and this can be due to several factors. Firstly, the high dimensionality of the data as well as the data structure could cause each model with their differing optimal hyperparameters to select different features. Particularly in the case of the XGBoost and elastic net models, where categorical features were one hot encoded resulting in 1,414 total predictors. Secondly, this could be due to the high degree of multicollinearity in the data. Time dependent features were included for each year of the time span of the data, which implies high correlation between features at different lags. In addition to conventionally correlated features, i.e. education and personal income, there is likely a degree of correlation between personal characteristics as well as neighborhood characteristics. Though the three methods used are relatively resistant to the negative effects of multicollinearity on prediction, this does render the interpretation of causal relationships difficult given the lack of consensus on feature importance.

Overall, prediction quality did not vary strongly with the method, though subtle differences were observed depending on the metrics and optimal risk thresholds. The XGBoost predictions of the out of sample validation data correctly classify 60.12% of movers 83.94% of stayers when choosing a risk threshold that minimizes the FPR within a range of risk thresholds that yield a TPR of between 60-80%. When taking the median threshold, the proportion of correctly classified stayers drops to 76.78%, while the proportion of correctly classified movers increases to 71.56%, achieving a balanced accuracy of 74.17%.

When applying derived risk thresholds to model predictions one year into the future, the XGBoost model correctly classifies 62.94% of movers correctly and 75.56% of stayers correctly, yielding a balanced accuracy of 69.25%. When optimizing for the maximal proportion of correctly classified movers and minimal proportion of incorrectly classified movers, the elastic net results in a TPR of 60.47% and an FPR of 22.77%. Though the aforementioned Dutch study predicted moving behavior within two years of a given reference date as opposed to one year as was done in this study, similar prediction quality was attained (TPR =60% and FPR=19%) when maximizing for the difference between the TPR and FPR. To this end, the present study prevailed in predicting moving behavior using public register data, though there are numerous opportunities to improve model accuracy and predictive performance.

5.2 Limitations

Though ROSE was ultimately chosen over random undersampling without replacement and random oversampling with replacement, the Synthetic Minority Over-sampling TEchnique (SMOTE), a widely used oversampling technique creating new synthetic minority class examples [38], could also be tried to rectify the unequal class distributions in the training data. Another technique to explore is class weights, that introduce a greater penalization for misclassification of the minority class. Additionally, rather than base the selection of the sampling technique on classification decision trees, it would be more optimal to train each of the models employed in the study (elastic net, random forest, XGBoost). This of course could be done in future research with sufficient time.

Expanding the time horizon as well as the study sample size are additional opportunities to improve model accuracy and reduce generalization error. In the Dutch study, 17 years of life history events and time dependent features were used to predict moving behavior, whereas in the present study only 4 years were used for prediction. Due to limitations on the time necessary for the appropriate data preprocessing and linking of public registers to expand the time horizon, only 4 years of life history events were included. Additionally, the measurement of certain important features, such as educational status, differed between years, which limited the possible time horizon.

Respecting the data privacy regulations and confidentiality of the public register data provided by Statistics Flanders, any data manipulation, preprocessing, and analyses was only possible on an on-site secure government server, on a provided computer. Given the limited processing power (8GB RAM) available, significant compromises were made for the sample size of the training data. Access to greater processing power as well as a remote secure server would permit increasing sample size, and reduce overfitting.

Expanding the set of features would be advantageous for predictive performance. Characteristics related to home ownership, for example, if someone was a homeowner or a renter, possessed high feature importance in Statistics Netherlands study on moving propensity. Migration status as well as migration type (which has been shown to have an effect on moving propensity) would permit expanding the study beyond Belgian citizens. Additionally, the inclusion of household type by income (private, student household with income), as well as student status could improve model performance, as they are likely tied with moving behavior. The aforementioned features were unavailable at the time of this study, but would be valuable additions to future work.

In this study, two way interactions between features were not included for the eNET model, due to the computational effort required for such high dimensional data following one-hot encoding (1,414 features). This presents a missed opportunity for a more representative model for moving propensity. The high dimensionality was as mentioned a result of the categorization strategy of features, as well as the inclusion of time dependent features for each year in the study. An alternative, could be to keep many of the categorized features as continuous (in cases where applicable), as well as only include the lags of certain features (i.e. life history events), while retaining time since variables as well as personal characteristics, as well as the cumulative sum for select life history events. In future work, it can be assessed if such changes result in better prediction quality when including two-way interactions.

In the case of the tree based models used in the study, namely random forest and XGBoost, many argue that both machine learning algorithms capture interactions. This is due to the recursive structure of decision trees where dependencies are accounted for hierarchically. Moreover, different behavior in two branches after a split implies potential interactions between predictors. [39] These interactions often get lost, as they are not captured in feature importance plots, which only indicate individual feature importance. Future research can explore potential 2-way interactions through the computation of Friedman's H-statistic, for example.

Considerable gains in predictive performance can be seen with the inclusion of a set of features related to social distance, as formulated in a previous study on moving propensity and social distance. [40] Given that neighborhood level features had been generated for the study, it would be feasible to engineer for a set of demographic characteristics a "social distance" feature. For example, if someone's income was much lower or much higher than the neighborhood median income. These features would likely have a higher impact on moving propensity than potential interactions captured in the tree based models between individual and neighborhood level characteristics.

5.3 Conclusion and Future Work

Though there is certainly room for improvement in terms of the methodology used in the present study to obtain better prediction quality, results point to the viability of the application secondary data for research in lieu of or at least complementary survey questionnaires.Moreover, such methodology can be applied to other research questions that can ultimately inform public policy. For instance, machine learning methods could be applied to public register data to predict the propensity to become unemployed, or require government financial assistance. In an increasingly digitized age of automated data collection, the richness and utility of secondary data sources only grows, creating more opportunities to harness the immense research potential of administrative records. With a gradual shift from survey questionnaires already in motion, the present research represents a valuable contribution to this trend.

6. References

1. Davies, R. S. (2020).’Survey Error and Response Bias Problems’ in *Designing Surveys for Evaluations and Research*. “ EdTech Books. https://edtechbooks.org/designing_surveys
2. Auerbach, J., Brummet, Q., Czajka, J., Hough, G. C., Hunsinger, E., & Salvo, J. (2019). Will administrative data save government surveys?. *Significance*, 16(5), 35-39.
3. Sturgis, P., & Luff, R. (2021). The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015. *International Journal of Social Research Methodology*, 24(6), 691-696.
4. Kessler, R. C., Bauer, M. S., Bishop, T. M., Demler, O. V., Dobscha, S. K., Gildea, S. M., ... & Bossarte, R. M. (2020). Using administrative data to predict suicide after psychiatric hospitalization in the veterans health administration system. *Frontiers in psychiatry*, 11, 390.
5. Garcia-Carretero, R., Roncal-Gomez, J., Rodriguez-Manzano, P., & Vazquez-Gomez, O. (2022). Identification and predictive value of risk factors for mortality due to Listeria monocytogenes infection: use of machine learning with a nationwide administrative data set. *Bacteria*, 1(1), 12-32.
6. Rosellini, A. J., Monahan, J., Street, A. E., Heeringa, S. G., Hill, E. D., Petukhova, M., ... & Kessler, R. C. (2016). Predicting non-familial major physical violent crime perpetration in the US Army from administrative data. *Psychological medicine*, 46(2), 303-316.
7. Burger, J., Buelens, B., de Jong, T., Gootzen, Y., Janssen-Jansen, S., van Beuningen, J., (2018), Een enquêtevraag vervangen door een modelvoorspelling op basis van registerdata, CBS Heerlen
8. de Jong, P. A. (2022). Later-Life Migration in The Netherlands: Propensity to Move and Residential Mobility. *Journal of Aging and Environment*, 36(1), 16-39.
9. Sonzogno, G. V., Urso, G., & Faggian, A. (2022). Migration propensity of peripheral youth: insights from Italy. *Regional Studies, Regional Science*, 9(1), 709-726.
10. Clark, W., Deurloo, M., & Dieleman, F. (2006). Residential mobility and neighbourhood outcomes. *Housing studies*, 21(3), 323-342.
11. Van der Vlist, A. J., Gorter, C., Nijkamp, P., & Rietveld, P. (2002). Residential mobility and local housing-market differences. *Environment and Planning A*, 34(7), 1147-1164.
12. Zorlu, A. (2009). Ethnic differences in spatial mobility: The impact of family ties. *Population, Space and Place,* 15(4), 323-342.
13. Mulder, C. H., & Hooimeijer, P. (1999). Residential relocations in the life course. In *Population issues* (pp. 159-186). Springer, Dordrecht.
14. Böheim, R., & Taylor, M. P. (2002). Tied down or room to move? Investigating the relationships between housing tenure, employment status and residential mobility in Britain. *Scottish Journal of Political Economy*, 49(4), 369-392.
15. De Groot, C., Mulder, C. H., Das, M., & Manting, D. (2011). Life events and the gap between intention to move and actual mobility. *Environment and planning A*, *43*(1), 48-66.

16. Musterd, S., Van Gent, W. P., Das, M., & Latten, J. (2016). Adaptive behaviour in urban space: Residential mobility in response to social distance. *Urban Studies*, 53(2), 227-246.
17. van Ham, Maarten & Clark, William. (2009). Neighbourhood Mobility in Context: Household Moves and Changing Neighbourhoods in the Netherlands. *Environment and Planning A*. 41. 1442-1459. 10.1068/a4182.
18. Fischer, Peter & Malmberg, Gunnar. (2001). Settled People Don't Move: On Life Course and (Im-)Mobility in Sweden. *International Journal of Population Geography*. 7. 357 - 371. 10.1002/ijpg.230.
19. Darlington-Pollock, F., Lomax, N., & Norman, P. (2019). Ethnic internal migration: The importance of age and migrant status. *The Geographical Journal*, *185*(1), 68-81.
20. Burger, J., Buelens, B., de Jong, T., Gootzen, Y., Janssen-Jansen, S., van Beuningen, J., (2018), Een enquêtevraag vervangen door een modelvoorspelling op basis van registerdata
21. *SILC definitions | Statbel*. (n.d.). https://statbel.fgov.be/en/themes/households/poverty-and-livingconditions/plus
22. Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*(2013), 332-340.
23. Devi, D., Biswas, S. K., & Purkayastha, B. (2020, July). A review on solution to class imbalance problem: undersampling approaches. In *2020 international conference on computational performance evaluation (ComPE)* (pp. 626-631). IEEE.
24. Demir, S. & Şahin, E. K. (2022). Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes . Avrupa Bilim ve Teknoloji Dergisi , Ejosat Special Issue 2022 (ICAENS-1) , 142-147 . DOI: 10.31590/ejosat.1077867
25. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*(1), 92-122.
26. Lunardon N, Menardi G, Torelli N (2014). "ROSE: a Package for Binary Imbalanced Learning." *R Journal*, 6(1), 82–92.
27. Hastie, T. (2009). (Second Edition, pp. 61-68, 73) [Review of *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*]. 233 Spring Street, New York, NY 10013, USA: Springer Series in Statistics.
28. Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.
29. Hastie, T. (2009). (Second Edition, pp. 587–588) [Review of *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*]. 233 Spring Street, New York, NY 10013, USA: Springer Series in Statistics.
30. Hastie, T. (2009). (Second Edition, pp. 309) [Review of *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*]. 233 Spring Street, New York, NY 10013, USA: Springer Series in Statistics.
31. *Introduction to Boosted Trees — xgboost 1.7.3 documentation*. (n.d.). https://xgboost.readthedocs.io/en/stable/tutorials/model.html

32. Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6), 15501329221106935.
33. Hastie, T. (2009). (Second Edition, pp. 367–368) [Review of *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*]. 233 Spring Street, New York, NY 10013, USA: Springer Series in Statistics.
34. Kuhn M, Wickham H (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*. https://www.tidymodels.org.
35. Friedman J, Hastie T, Tibshirani R (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software*, 33(1), 1–22. doi:10.18637/jss.v033.i01,
36. Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” *R News*, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.
37. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
38. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
39. Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests?. *BMC bioinformatics*, *17*(1), 1-10.
40. Musterd, S., Van Gent, W. P., Das, M., & Latten, J. (2016). Adaptive behaviour in urban space: Residential mobility in response to social distance. *Urban Studies*, *53*(2), 227-246.

AFDELING
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be