DETECTING FACTORS AFFECTING CONTRACT TERMINATIONS IN THE ELECTRICITY DISTRIBUTION SYSTEM

Lappeenranta-Lahti University of Technology LUT

Master's Program in Computational Engineering, Master's Thesis

2022

Markku Karhunen

Examiner: Professor Heikki Haario
Professor Jukka Lassila

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering

Markku Karhunen

Detecting factors affecting contract terminations in the electricity distribution system

Master's thesis
2022
42 pages, 9 figures, 8 tables, 2 appendices
Examiners: Professor Heikki Haario and Professor Jukka Lassila

Keywords: Distribution system operators (DSOs), contract terminations, geolocation, machine learning

Distribution system contract terminations are a factor of economic interest in the electricity market, as they affect the willingness of the distribution system operators (DSOs) to invest in the improvement of the distribution grid, such as installation of underground cables. In this study, statistical and machine learning methods were employed to study factors affecting the distribution system contract terminations. To this end, a data set obtained from four DSOs was used. Each contract was geolocated and combined with data obtained from government authorities regarding the socioeconomic situation in the area. The final data set involved 434 terminated contracts, 41,104 non-terminated contracts and 99 variables. Subsequently, regression analysis was used to study the effects of different covariates. Multivariate model choice was performed and two machine learning models were trained to predict the termination status. It was found that positive net migration and high local income level decreased the risk of contract terminations, whereas high values of the age of the contract holder, distance to a city and distance to a lake increased the risk. The models constructed in this study could not predict the termination status as a binary yes/no variable. However, it was possible to use the models to calculate risk scores for individual customers. For example, for a multivariate logit model, values as high as 1.9% were frequently observed (whereas the average risk was 1.0%). This may be compared to credit risk analysis where similar low risk scores of a few per cents are often encountered.

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Laskennallinen tekniikka

Markku Karhunen

Tekijöitä jotka vaikuttavat sähköverkkosopimusten irtisanomiseen

Diplomityö
2022
42 sivua, 9 kuvaa, 8 taulukkoa, 2 liitettä
Tarkastajat: Professor Heikki Haario ja Professor Jukka Lassila

Hakusanat: Jakeluverkkoyhtiöt, sopimusten irtisanominen, geolokaatio, koneoppiminen
Keywords: Distribution system operators (DSOs), contract terminations, geolocation, machine learning

Jakeluverkkosopimusten irtisanominen on taloudellisesti merkittävä ilmiö sähkömarkkinoilla, koska se vaikuttaa jakeluverkon haltijoiden haluun investoida verkon parantamiseen esim. asentamalla maakaapeleita. Tässä työssä tutkittiin näihin irtisanomisiin vaikuttavia tekijöitä tilastollisten ja koneoppimismenetelmien avulla. Työssä käytettiin neljältä eri jakeluverkkoyhtiöltä saatua aineistoa. Jokainen jakeluverkkosopimus kohdennettiin maantieteelliseen koordinaattijärjestelmään ja yhdistettiin viranomaisilta saatuun, alueen sosioekonomisia oloja koskevaan tilastoaineistoon. Lopullinen aineisto käsitti 434 irtisanottua sopimusta, 41 104 ei-irtisanottua sopimusta ja 99 muuttujaa. Regressioanalyysia käytettiin eri muuttujien vaikutuksen selvittämiseen. Työssä käytettiin monimuuttujamalleja ja kaksi koneoppimismallia opetettiin ennustamaan sopimusten irtisanomista. Havaittiin että muuttovoitto ja korkea paikallinen tulotaso vähensivät sopimusten irtisanomisen riskiä, kun taas sopimuksen haltijan korkea ikä, etäisyys maakuntakeskukseen ja etäisyys järven rantaan lisäsivät riskiä. Tässä työssä rakennetut mallit eivät pystyneet ennustamaan sopimusten irtisanomista binäärisenä kyllä/ei-muuttujana. Mallien avulla pystyttiin kuitenkin laskemaan asiakaskohtaisia risk scoreja. Esimerkiksi logit-monimuuttujamalli tuotti useille asiakkaille 1,9 % risk scoren (siinä missä keskimääräinen riski oli 1,0 %). Tätä voidaan verrata luottoriskianalyysiin, jossa myös usein esiintyy vastaavia matalia, parin prosentin risk scoreja.

ACKNOWLEDGEMENTS

I thank Prof. Heikki Haario and Prof. Jukka Lassila for supervising my thesis. I also thank Prof. Haario for giving me the Matlab code used for the stepwise OLS analysis. I thank the Laboratory of Electricity Market and Power Systems for providing me with the data and the research question and for offering me a summer job.

I thank my parents for supporting my studies in my early middle age.

Lappeenranta, March 24, 2022

Markku Karhunen

LIST OF ABBREVIATIONS

AIC	Akaike's information criterion
AUC	area under curve
DSO	distribution system operator
E_j	marginal effect of variable j
NSLM	non-standard logit machine
OLS	ordinary least squares
OOS	out-of-sample, e.g. performance in validation data
P	P value of classical statistics
P_B	Bonferroni-corrected P value
P^*	Bonferroni-corrected significance threshold
p^*	cut-off for probability scores, not to be confused with P^*
PC	principal component
Q^2	OOS coefficient of determination
r	Pearson correlation coefficient
ROC	receiver operating characteristic
RSS	residual sum of squares
*	effect significant at 0.05 level
**	effect significant at 0.01 level
***	effect significant at 0.001 level

CONTENTS

1 Introduction **8**
1.1 Background 8
1.2 Objectives and delimitations 9
1.3 Structure of the thesis 10

2 Related work **11**

3 Material and methods **13**
3.1 Empirical data 13
3.2 Estimates of migration 14
3.3 Normalisation 16
3.4 Statistical modelling vs. machine learning 16
3.5 Logistic regression 17
3.5.1 Definition 17
3.5.2 Multiple testing correction 19
3.6 Multivariate model choice 20
3.6.1 AIC minimisation 21
3.6.2 Stepwise OLS 21
3.7 Non-standard logit machine 22
3.7.1 Definition 22
3.7.2 Testing 23

4 Results **26**
4.1 Univariate analyses 26
4.2 Multivariate analyses 28
4.3 Non-standard logit machine 30
4.4 Sensitivity analysis 32

5 Discussion **36**
5.1 Current study 36
5.2 Future study 37

6 Conclusion **39**

REFERENCES **40**

APPENDICES
Appendix 1: Inclusion criteria.
Appendix 2: Data sources.

1 Introduction

1.1 Background

In many countries, the electricity market is organised so that there is a distinction between suppliers and distribution system operators (DSOs). The suppliers are the energy producers, whereas the DSOs control the distribution network (cables and overhead lines), being typically local monopolies. The DSOs' operating environment has changed in the past decades, with society being increasingly reliant on a secure supply of electric energy. Globally, the distribution systems are subject to emerging trends, among them distributed energy resources, solar generation and electrification of transportation and heating [1–3]. All these lead to increased expectations of the power quality, security of supply and efficiency of the distribution network. Smart grids and peer-to-peer trading have been proposed as solutions to these challenges [3–5]. Furthermore, it is also possible for a customer to terminate a DSO contract when the need of electricity ends or the customer switches to off-grid local generation which complicates the planning of the power system.

In this thesis, the focus is on DSO contract terminations as these represent an economic risk for a DSO planning to improve the security of power supply. The thesis uses big-data techniques to disentangle factors which are associated with the DSO contract terminations in the electricity market of the rural Eastern Finland.

In Finland, the DSOs are trapped in a legal trilemma: 1. The law requires high security of supply from the distribution network, limiting the maximum allowable duration of electricity supply interruptions; 2. the Energy Authority regulates the unit prices of the network components and the reasonable rate of return of the DSOs, while 3. the DSOs must be economically profitable, being mostly private enterprises [6]. The most straightforward way to improve the security of supply would be to replace the overhead lines by underground cables, but this, in turn, requires investments which are limited by the profitability of the DSOs.

The cost of underground cables is typically significantly higher than that of overhead lines. This may be a challenge, especially in sparsely populated areas. The overall population density of Finland is 16 inhabitants/km^2, whereas in rural areas in Eastern Finland, the population density ranges typically from 0 to 12 inhabitants/km^2. This implies that the cable length per individual customer can be quite considerable. In these circumstances, the investment decision of the DSO can be characterised by two main options: A. install

the underground cables and take the risk that the contracts are terminated, or B. delay grid infrastructure investments and be prepared to repair any faults in the required time window, which also creates costs. If the DSO knows that the contracts concerned are likely to be terminated, it shifts the balance in favour of option B. Thus, there has been recent interest towards contract terminations from the part of DSOs [6].

1.2 Objectives and delimitations

The objectives of this thesis are three-fold:

1. To detect factors which are associated with the contract terminations.
2. To validate these associations in a geographically distinct area.
3. To construct a machine learning model to predict the termination status of individual contracts.

There are a number of limitations to this work. Most importantly, the DSO contract termination is a relatively rare event, with 1.0% of customers terminating their contract in these data. This implies that predicting the termination status is difficult, or rather, that it is difficult to improve the prediction upon the baseline guess "does not terminate". On the other hand, even if prediction is not possible, statistically significant associations between the contract terminations and different covariates may be found.

Furthermore, there are two major technical limitations: Firstly, most covariates were measured in 5 km $\times$ 5 km grid cells, whereas each grid cell contains many DSO contracts. This is bound to affect the accuracy of the predictions, but because of the privacy regulation, it could not be alleviated at this point. (Of course, some covariates, such as number of jobs or population, only have an areal interpretation and do not refer to individual contracts, as such.) Secondly, official statistics do not exist for the net migration and net population change in the grid cells. However, this thesis presents a method to calculate estimates of these quantities, as they were deemed to be important covariates (Subsection 3.2).

1.3 Structure of the thesis

Section 2 presents a short summary of the previous literature relevant for studying DSO contract terminations. Section 3 introduces the empirical data and the statistical methods. Section 3 is divided so that Subsection 3.1 presents a verbal description of the data and Subsections 3.2-3.3 introduce the straight-forward calculations applied on the data, whereas Subsections 3.4-3.7 are more theoretical, introducing the theory behind the models used in this study. Section 4 presents the results. The results from training data and validation data are discussed side by side. Section 5 presents a discussion and Section 6 concludes. Appendix 1 presents the inclusion criteria used in this study and Appendix 2 presents a list of the data sources.

2 Related work

The DSO contract terminations are a subject which has been studied very little. To my best knowledge, the only direct precursor of this work is the Master's thesis of Arimo Perosvuo [6] which used largely the same data as this study. Perosvuo found there to be weak predictability in the contract terminations. He also identified associations between the contract terminations and a few covariates. Distance to a lake was the most important among these variables. This results from the fact that the proximity of a lake or sea is a desired property for a Finnish residential building, and thus, the DSO contracts are less likely to be terminated from buildings near the lakes in Eastern Finland.

Perosvuo also developed a predictive model for the contract terminations. This was based on a stratification of data. An example illustrates the method of Perosvuo: If a DSO was interested in the liability of over 70-year-olds living 2 km away from a lake to terminate their contracts, it would calculate the termination probability as

$$\hat{P} = \frac{B}{A + B - C} \tag{1}$$

where B is the number of terminated contracts in the said customer segment, A is the number of current customers in that segment and C is the number of new customers entering the data during the follow-up period. However, as Perosvuo noted, the mechanistic application of Eq. 1 is problematic for a rare event such as a DSO contract termination. Namely, if the DSO does not observe any terminations in some small customer segment, Eq. 1 will give $\hat{P} = 0$ which is not a sensible estimate. To this end, it is desirable to use regression models to analyse the effects of different covariates on the termination probability. In regression models, it is possible to specify the effects of different covariates as continuous functions which increases the degree of realism and also enables one to avoid the unrealistic predictions of $\hat{P} = 0$ and $\hat{P} = 1$.

This work differs from Perosvuo's analysis [6] in a number of ways: 1. This thesis uses multivariate regression analysis as a modelling technique and introduces a new machine learning method, the non-standard logit machine, to predict the termination status. 2. In this work, net migration and net population change in the area are used as predictive factors. 3. This thesis divides the data in training and validation data sets which helps to establish the statistical validity of the results.

Apart from the thesis work of Perosvuo [6], one may speculate more widely on the factors potentially affecting the contract terminations. It can be expected that geographical factors

partially explain the pattern: Contracts are usually terminated from buildings which are abandoned or left empty with weak prospects of getting the property sold. In Finland, such buildings are found primarily in areas with a negative migration rate [6]. Thus, it may be expected that the migration rate is associated with the contract terminations. Large areas of Eastern Finland have a negative net migration rate, resulting from urbanisation [7].

The factors affecting urbanisation and the pattern of migration in Finland have been studied to some degree, with unemployment identified as a predictive factor for negative net migration [8]. It is also known that the migration flows are spatially highly aggregated, especially among people with higher education [9] and they seem to form self-sustaining cycles [10]. It is also known that there are health differences in migration, with healthier individuals moving longer distances [11].

Generally, the pattern is such that the population concentrates in the metropolitan area of Helsinki and in a few regional cities, while the rural areas are being depopulated [7]. However, the seasonal population is an exception to this rule [12]. The seasonal population is represented by holiday homes and second homes in the distribution system data; dwellings of these kinds are increasingly popular in the present-day Finland. As a result of these considerations, this thesis specifically concerns data from small residential buildings in rural areas, excluding holiday homes. Data from four DSOs are used and these are integrated with register-based data from government authorities, subject to privacy regulation.

3 Material and methods

3.1 Empirical data

The data originated from four Finnish DSOs, comprising tens of thousands of contracts. The contract terminations had occurred between 16 Sep. 2013 and 27 Oct. 2019. They were coded as a binary response variable and contrasted against geographic and demographic covariates collected from the year 2010. There were a few inclusion criteria applied to individual contracts. For example, it was required that all contracts had a fairly similar follow-up time. A panel of experts, i.e. electricity market scholars, was used to define these inclusion criteria. For more details, see Appendix 1. There were a total of 41,538 DSO contracts included in the final sample.

The source of most geographical covariates was the Grid Database from Statistics Finland. However, data from the National Land Survey of Finland and the Finnish Environment Institute were also used. (See Appendix 2 for a detailed list of data sources.) The data originating from Statistics Finland were measured in 5 km $\times$ 5 km grid cells, not to be confused with the electric grid. These variables were divided into six main categories: demography, education, income, households, buildings and jobs.

Apart from the geographical variables, three individual-based variables were also calculated for each contract: age of the contract holder (from the DSO data), and distance to a lake and distance to a city which were based on the coordinates of each individual building. For calculating the distance to a city, the main regional cities were used: Kuopio for North Savo, Mikkeli for South Savo, Kouvola for Kymenlaakso and Joensuu for North Karelia. The geolocation was performed in QGIS, while downstream analyses were performed in R, MATLAB and Python. Most of the analyses were carried out in R, whereas MATLAB was used to run the algorithms crosback and crosforw [13] and Python was used to implement the neural networks (see Subsection 3.7.2).

The data were divided so that the data from North Savo, South Savo and Kymenlaakso were used as a training data set, comprising 358 terminated contracts and 27,197 non-terminated contracts, whereas the data from North Karelia were used as a validation data set, comprising 76 terminated contracts and 13,907 non-terminated contracts. However, a sensitivity analysis was run so that North Savo was used as the validation data set (see Subsection 4.4).

To conclude the discussion on the empirical data, Fig. 1 presents the data on a map. Fig. 2 presents the workflow in a schematic form.

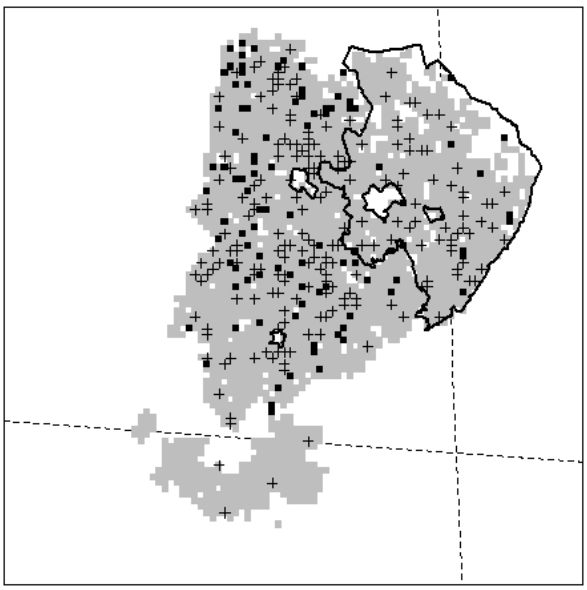

Figure 1. Map of the raw data: The geographic data were available from the grey area. The crosses are grid cells where contract terminations have taken place, and the black boxes are grid cells where over 10% of the contracts were terminated. The large area on the right represents the validation data. The dashed lines are 61°N and 30°E.

3.2 Estimates of migration

Data from Statistics Finland were used to calculate the approximate number of migrants in each grid cell. The initial age distribution of each grid cell in 2010 was divided into one-year age groups. Subsequently, a recursive procedure was carried out for these age distributions. In more detail, the age cohort mortality of 2010 (Statistics Finland) was used to calculate the expected number of persons in each age group, each grid cell by using the formula

$$\hat{n}_{i+1,j+1} = \hat{n}_{i,j}(1 - \mu_i) \tag{2}$$

where μ_i denotes the age cohort mortality in age group i, with $i = 7, \ldots, 94$ and $j = 2010, \ldots, 2017$. Thus, $\hat{n}_{i,j}$ measures the expected age distribution in absence of migra-

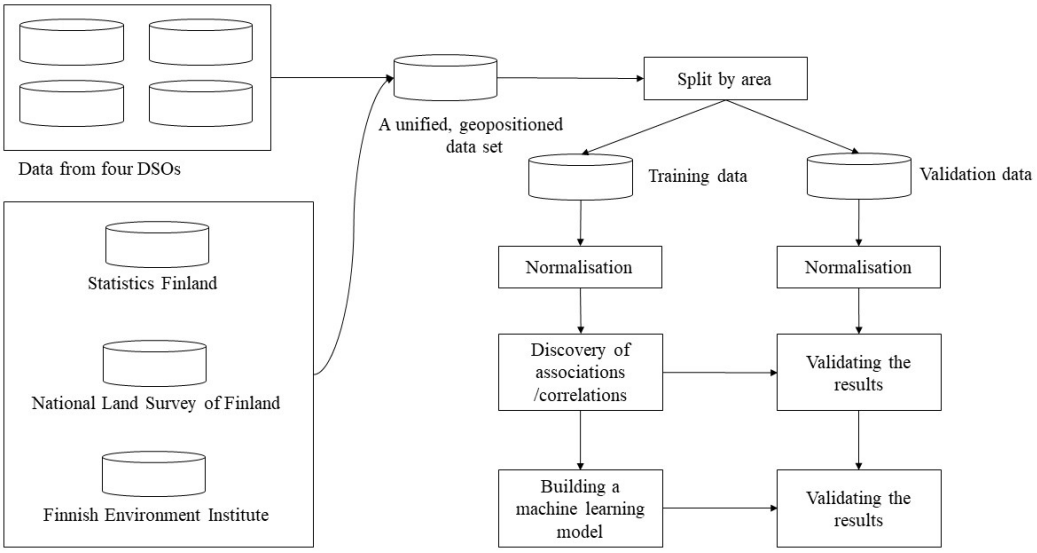

Figure 2. Block diagram of the data flow: Training data represents the majority of the data, i.e. North Savo, South Savo and Kymenlaakso. Validation data contains North Karelia. Discovery of associations involves logistic regression models, whereas the machine learning method is the new model introduced in this thesis, the non-standard logit machine.

tion. The inferred net migration was calculated from this recursion by using

$$m = \sum_{i=8}^{94} n(i, 2018) - \sum_{i=8}^{94} \hat{n}(i, 2018) \tag{3}$$

where $n(i, j)$ denotes the observed number of individuals of age i on year j. The net change of population could be calculated from the observed data simply as

$$\Delta n = \sum_{i=0}^{\infty} n(i, 2018) - \sum_{i=0}^{\infty} n(i, 2010). \tag{4}$$

Notably, Δn is directly observed but it concerns the effects of all demographic forces combined, i.e. migration, mortality and birth rate. The calculation of m was motivated by the fact that it measures the effect of migration. The effects of mortality and birth rate could be deducted away by applying the recursion and by considering a population of sufficient age.

3.3 Normalisation

Upon visual inspection of the training data, most variables were highly non-normal, thus rendering them unreliable to be used as covariates in a regression analysis. Thus, these variables were normalised by using the transformation

$$f(x) = \text{sign}(x)\text{lg}(|x| + 1). \tag{5}$$

Some variables had a naturally constrained range of variation and moderate or mild non-normality. These variables were not transformed. The variables which were not transformed were age, distance to a city, median income and median income of households.

3.4 Statistical modelling vs. machine learning

This study uses two types of models. The logistic and linear regression models can be termed as classical statistical models, whereas the non-standard logit machine and its benchmark model, the neural network, can be termed as machine learning models. Machine learning models are typically used to construct predictions of a target variable (such as contract termination) based on different covariates (such as contract holder age, distance to a lake, etc.).

Regression models can also be used for prediction in cases where the signal in the data is sufficiently strong. However, they are often used to detect mere 'associations' between the outcome and the variables of interest, the covariates. The customary criterion of detecting an association is the statistical significance of the relevant regression coefficient. It is important to note that a model may have highly significant regression coefficients without any predictive power.

> *Illustrative example:* Consider an epidemiological study where the research team measures the effect of sunlight exposure and ice cream consumption on the incidence of skin cancer. (Assume that all of these can be reliably measured.) It turns out that the sunlight exposure is associated with the skin cancer with $P = 0.002$, i.e. it is highly significant. However, based on these variables, it is hardly possible to predict the onset of skin cancer for any particular individual. For example, the case may be that the life-time risk of developing a skin cancer is 15% and the maximal dose of sunlight observed in the data increases the risk by 200%. Thus, for this individual, the predicted

probability of skin cancer is 45%, i.e. still below 50%. Arguably, the present study with the DSO contract terminations may have similar problems, as only 1.0% of the contracts were terminated.

Indeed, the regression coefficients are parameters, i.e. something internal to the model, and their statistical significance does not guarantee that the model predicts the data very well, or vice versa. However, as the purpose of the machine learning models is to predict the data, their predictive power needs to be tested. To avoid over-fitting, it is important to assess specifically the out-of-sample (OOS) predictive power. To this end, either cross validation or separate test data (also known as hold-out data) can be used. In this thesis, separate test data were used because this strategy enables one to validate the associations of the statistical models.

The use of separate test data is common in statistical genetics where weak associations are sought in geographically heterogeneous populations. In more detail, König [14] makes a distinction between replication and validation: An effect can be considered replicated if the significance can be re-established in the same study population (e.g. North Savo), whereas it can be considered validated if the significance can be established in a new population (e.g. North Karelia as opposed to North Savo). Successful validation seems to indicate that the results of a study are something substantial, not just statistical artefacts. Thus, this approach was chosen for this study.

3.5 Logistic regression

3.5.1 Definition

Logistic aka. logit regression is a generalized linear model specifically suited for binary data [15]. In logistic regression, the probability that observation i is positive (e.g. a termination) is defined as

$$p_i = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \tag{6}$$

where $\boldsymbol{x}_i$ is the vector of covariates for observation i and $\boldsymbol{\beta}$ is the vector of regression coefficients. In the univariate regressions of this thesis, $\boldsymbol{x}_i = (1, x_{ij})$ and $\boldsymbol{\beta} = (\beta_0, \beta_j)$ so that the model involves a constant term β_0. In the multivariate regressions, a constant term is similarly involved, and consequently, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{id})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)$.

In logit regression, $\boldsymbol{\beta}$ is typically estimated by using the maximum-likelihood principle.

The likelihood function in the logit regression is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{7}$$

where $y_i = 0, 1$ is the observed label. (In passing, it should be mentioned that the logit regression can be defined so that each y_i involves a binomial trial, instead of a binary one [15], but in this thesis, the data are binary.) Consequently, the log-likelihood is

$$\ell = \sum_{i=1}^{n} \Big(y_i log(p_i) + (1-y_i) log(1-p_i) \Big) \tag{8}$$

and the first-order condition for the maximum likelihood is

$$\nabla \ell = \sum_{i=1}^{n} \Big(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \Big) \nabla p_i = \mathbf{0} \tag{9}$$

where

$$(\nabla p_i)_j = \frac{\partial p_i}{\partial \beta_j} = p_i(1-p_i)x_{ij}. \tag{10}$$

Alas, Eq. 9 needs to be solved numerically. An algorithm known as iteratively reweighted least squares is typically used [15, 16]. As usual with maximum-likelihood models, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained from the observed information matrix

$$\boldsymbol{J}(\hat{\boldsymbol{\beta}}) = -\Big[\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} \Big]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \tag{11}$$

where $\hat{\boldsymbol{\beta}}$ denotes the maximum-likelihood estimate. Asymptotically,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{J}(\hat{\boldsymbol{\beta}})^{-1}). \tag{12}$$

The confidence intervals and P values can be obtained from the marginal distributions implied by Eq. 12. In case of logit regression, $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$ gets the form

$$\boldsymbol{J}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}' \mathrm{diag}\{p_i(1-p_i)\} \boldsymbol{X}, \tag{13}$$

as noted by McCullagh and Nelder [15]. In practice, the software [16] calculates the P value of variable j as

$$P = 2\Phi\Big(\frac{-|\hat{\beta}_j|}{\mathrm{SD}(\hat{\beta}_j)} \Big) \tag{14}$$

where Φ is the cumulative distribution function of the standard normal distribution and $\mathrm{SD}(\hat{\beta}_j)$ is the standard deviation of $\hat{\beta}_j$ implied by Eq. 12. This is a two-sided P value

based on the normal approximation.

Implicitly, it is thought that each P value corresponds to a statistical test, i.e. the investigator testing whether e.g. $P < 0.05$ or $P < 0.01$. Thus, if there are 100 P values in a study, there are 100 statistical tests. (The importance of this observation is discussed in Subsection 3.5.2.) In passing, it should be mentioned that the effect is of course more significant, if the P value is smaller. It is common to use $P < 0.05$ as a cut-off and that was applied in this thesis, subject to multiple testing correction (Subsection 3.5.2).

Finally, to illustrate the results, the marginal effects of the variables used in the logit models were calculated. In this thesis, the marginal effect of variable j was defined as

$$E_j = \frac{\partial p_i}{\partial x_{ij}} \mid_{\boldsymbol{x}_i = \bar{\boldsymbol{x}}} \sigma_j = p_i(1 - p_i) \mid_{\boldsymbol{x}_i = \bar{\boldsymbol{x}}} \beta_j \sigma_j, \tag{15}$$

with $\bar{\boldsymbol{x}}$ denoting the sample mean of $\boldsymbol{x}$ and σ_j denoting the sample standard deviation of variable j. Thus, E_j measures how much a 1SD change of variable j would change the probability of termination, other things being equal.

3.5.2 Multiple testing correction

Multiple testing correction means correcting the significance levels of the statistical tests for the number of tests in the study, e.g. [17]. The significance level is typically measured by the P value. The P value can be defined as the probability that the test would detect a significant association under the null hypothesis, i.e. under the assumption that there is no association between $\boldsymbol{x}_j$ and $\boldsymbol{y}$. Such events are unavoidable, and they occur more and more frequently, the more tests there are.

> *Illustrative example:* Consider an engineer using Internet of things to study the breakdown of a large industrial machine. Assume that the specific cause of the breakdown is not known, but the engineer can use the measurement data from 1,000 sensors and test which of these correlate with the breakdown. Assume that the engineer uses the rule $P < 0.05$ to detect significant correlations. Consequently, the probability of detecting a false positive association is 0.05 for each individual test. Consequently, there will be some $0.05 \times 1{,}000 = 50$ correlations declared statistically significant, *even if the sensor output is random noise*. It seems that the P values need to be corrected somehow.

Luckily, there are a number of methods to adjust the P values. The Benjamini-Hochberg procedure [18] is used quite frequently in contemporaneous scientific literature. However, the Bonferroni correction was chosen for this thesis because of its transparency and easy interpretation. Unlike the Benjamini-Hochberg correction, it can be calculated directly from the number of tests (N) and the P value of each individual test. The Bonferroni-corrected P value is defined as

$$P_B = \min\{PN, 1\}. \tag{16}$$

Then, what are such P_B values good for? Consider the case that there are no real effects in the data and $P < 0.05/N = P^*$, i.e. $P_B < 0.05$, is used to declare a test as a significant one. Thus, P^* is the probability of detecting a false positive association for each individual test j. Denote this event by A_j. The probability of detecting at least one false positive association is

$$\mathbb{P}\Big(\bigcup_{j=1}^{N} A_j\Big) \leq \sum_{j=1}^{N} \mathbb{P}(A_j) = NP^* = 0.05. \tag{17}$$

Thus, Bonferroni correction controls the rate of false positive discoveries, also known as Type I error rate in the statistical literature [17].

3.6 Multivariate model choice

The data involved many seemingly redundant features, such as average income of individuals vs. average income of households (see Appendix 2). Thus, it was seen as necessary to reduce the dimensionality. An approach based partially on expert opinion was chosen. In more detail, the covariates were grouped in six conceptual categories, i.e. demography, education, income, households, buildings and jobs. The individually measured covariates were left out of these groups. Then, the first principal component (PC) was calculated for each of these groups and the sign of that PC was chosen so that the PC had an intuitive interpretation, e.g. PC1 of demography correlating positively with the number of people in the grid cell. Subsequently, the individual covariates and the PCs thus obtained were used as building blocks for multivariate models. Thus, the number of covariates was reduced to nine (3 individual covariates + 6 principal components).

3.6.1 AIC minimisation

Akaike's information criterion (AIC) is often used in empirical research to choose between competing models [19]. For a regression model, it can be defined as

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2d \tag{18}$$

where $\ell(\hat{\boldsymbol{\beta}})$ is the maximised log-likelihood of the model and d is the number of parameters in that model. The model with minimal AIC is chosen. Intuitively, AIC merits the model for giving a high likelihood and penalises it for using a lot of parameters (the $2d$ term), thus inducing parsimony. AIC is supposed to select asymptotically, i.e. as $n \to \infty$, the model which is closest to the unknown true model [20].

AIC minimisation was performed as follows:

1. The model space was indexed by a nine-dimensional vector $\boldsymbol{\phi} \in \{0, 1\}^9$ with $\phi_j = 0$ indicating $\beta_j = 0$ and $\phi_j = 1$ indicating $\beta_j \neq 0$.
2. Models with all possible values of $\boldsymbol{\phi}$ were estimated by maximum likelihood and the AICs were calculated.

3.6.2 Stepwise OLS

Models based on ordinary least squares (OLS) were also constructed. These were based on the data analysis toolbox of Haario [13]. For each combination of variables $\boldsymbol{\phi}$, the OLS estimate was calculated as

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{19}$$

Subsequently, $\hat{\boldsymbol{\gamma}}$ was used to calculate predictions for new observations as

$$\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}. \tag{20}$$

For each combination of variables $\boldsymbol{\phi}$, a Q^2 statistic was calculated as a proxy of OOS predictive power. The Q^2 is defined as

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})} \tag{21}$$

where $\tilde{y}_i$ is the prediction of y_i based on the complement of $(y_i, \boldsymbol{x}_i)$ and $\bar{y}$ is the mean of $\boldsymbol{y}$. In more detail, $\tilde{y}_i$ is defined as

$$\tilde{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\gamma}}_{-i} = \boldsymbol{x}_i' (\boldsymbol{X}'_{-i} \boldsymbol{X}_{-i})^{-1} \boldsymbol{X}'_{-i} \boldsymbol{y}_{-i} \tag{22}$$

where $\hat{\boldsymbol{\gamma}}_{-i}$, $\boldsymbol{X}_{-i}$ and $\boldsymbol{y}_{-i}$ denote $\hat{\boldsymbol{\gamma}}$, $\boldsymbol{X}$ and $\boldsymbol{y}$ calculated by deleting row i from the data. Importantly, however, Q^2 can be calculated without fitting n regressions [21]: It can be shown that

$$\tilde{y}_i = y_i - \frac{y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\gamma}}}{1 - h_{ii}} \tag{23}$$

where h_{ii} is obtained from the ortho-projection matrix

$$\boldsymbol{H} = \boldsymbol{X} (\boldsymbol{X}' \boldsymbol{X})^{-1} \boldsymbol{X}'. \tag{24}$$

A quasi-global maximisation of Q^2 was performed by two methods, crosback and crosforw [13]. Both methods move along the coordinate axes of $\boldsymbol{\phi}$. However, crosback starts from $\phi_j \equiv 1$, i.e. from the full model, and removes one variable at a time, whereas crosforw starts from $\phi_j \equiv 0$, i.e. from the empty model, and adds one variable at a time.

3.7 Non-standard logit machine

3.7.1 Definition

Finally, a machine learning model was built for the contract terminations. It was desired that this method would be easily interpretable and informative regarding the predictive power of the different covariates. The non-standard logit machine (NSLM) was defined as

$$p_i = \theta_x \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_x)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_x)} + \theta_z \frac{\exp(\boldsymbol{z}_i' \boldsymbol{\beta}_z)}{1 + \exp(\boldsymbol{z}_i' \boldsymbol{\beta}_z)} + \theta_w \frac{\exp(\boldsymbol{w}_i' \boldsymbol{\beta}_w)}{1 + \exp(\boldsymbol{w}_i' \boldsymbol{\beta}_w)}, \tag{25}$$

where $\boldsymbol{x}_i$ denotes the three individually measured covariates, $\boldsymbol{z}_i$ denotes the local covariates calculated from the same grid cell and $\boldsymbol{w}_i$ denotes the same covariates as $\boldsymbol{z}_i$, but calculated from the *neighbouring cells*, as depicted in Fig. 3. θ_x, θ_z and θ_w are parameters to be calibrated.

In practice, $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$ involved the six PCs calculated from the geographical data. (Note that in Subsections 3.5 and 3.6, $\boldsymbol{x}_i$ denotes *all* covariates.) Effectively, NSLM combines predictions obtained from three different logit models, one individually specific (the $\boldsymbol{x}$

term), one local (the $\boldsymbol{z}$ term) and one areal (the $\boldsymbol{w}$ term). Ideally, one would require

$$\theta_x + \theta_z + \theta_w = 1, \tag{26}$$

so that $\theta_x, \theta_w, \theta_z \geq 0$, to constrain p_i in (0,1), as is logically valid. Ideally as well, one would use the maximum-likelihood principle to estimate jointly $(\boldsymbol{\theta}, \boldsymbol{\beta}_x, \boldsymbol{\beta}_z, \boldsymbol{\beta}_w)$. However, due to numerical difficulties, another approach was chosen for this thesis: Initially, usual logistic regressions were used to estimate each of $(\boldsymbol{\beta}_x, \boldsymbol{\beta}_z, \boldsymbol{\beta}_w)$. Subsequently, $\boldsymbol{\theta}$ was calibrated by using two complementary methods:

1. Exhaustive search on a grid of points on the simplex Δ^2
2. OLS regression of y_i against $(\hat{p}_i(\boldsymbol{X}), \hat{p}_i(\boldsymbol{Z}), \hat{p}_i(\boldsymbol{W}))$.

Above, $\hat{p}_i(\boldsymbol{M})$ denotes the logit prediction obtained from matrix $\boldsymbol{M}$. Thus, e.g.

$$\hat{p}_i(\boldsymbol{X}) = \frac{\exp(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_x)}{1 + \exp(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_x)} \tag{27}$$

where $\hat{\boldsymbol{\beta}}_x$ denotes the maximum-likelihood estimate. In more detail, in method 1, a quasi-optimal value of $\boldsymbol{\theta}$ was sought by using a K-fold cross validation. The cross-validation scheme was such that on each cross-validation fold, one 10^{th} of the training data was randomly sampled to be used as hold-out data A_k. Thus, for each value of $\boldsymbol{\theta}$, the residual sum of squares (RSS) was calculated as

$$RSS = \frac{10}{K} \sum_{k=1}^{K} \sum_{i \in A_k} (y_i - p_i(A_k^c))^2 \tag{28}$$

where $p_i(A_k^c)$ is the prediction obtained by estimating $(\boldsymbol{\beta}_x, \boldsymbol{\beta}_z, \boldsymbol{\beta}_w)$ in the complement of A_k. $K = 400$ was used. In method 2, the estimate of $\boldsymbol{\theta}$ was obtained as

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{P}}' \hat{\boldsymbol{P}})^{-1} \hat{\boldsymbol{P}}' \boldsymbol{y} \tag{29}$$

where $\hat{\boldsymbol{P}} = (\hat{\boldsymbol{p}}(\boldsymbol{X}) \;\; \hat{\boldsymbol{p}}(\boldsymbol{Z}) \;\; \hat{\boldsymbol{p}}(\boldsymbol{W}))$.

3.7.2 Testing

The OOS performance of machine learning models needs to be tested, to avoid overfitting and to assess the practical usefulness of the proposed methods. To this end, an OOS

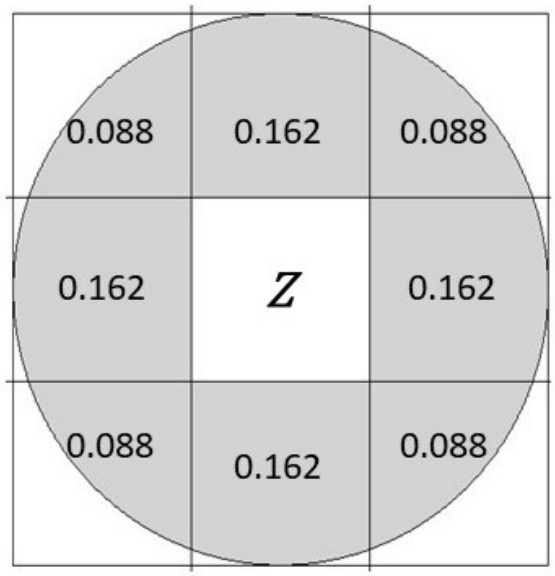

Figure 3. Weights used for calculating the areal covariates (the $\boldsymbol{W}$ matrix). These covariates were used in the machine learning models. The weights were chosen on basis of the overlap of each grid cell and a 7.5 km-radius circle. The area denoted by $\boldsymbol{Z}$, i.e. the contract's own grid cell, did not contribute to $\boldsymbol{W}$ but was used to calculate the local covariates (the $\boldsymbol{Z}$ matrix).

prediction exercise was carried out with the NSLM as depicted in Fig. 2. Two methods were used to measure the predictive power: the receiver operating characteristics (ROC, [22]) and confusion matrices.

ROC can be explained as follows: The binary classification methods typically give a probability score p_i for each observation i. This is, in turn, truncated by using some cut-off $p^* \in (0, 1)$ to give a binary prediction $\hat{y}_i \in \{0, 1\}$. (This is also the case for NSLM.) The sensitivity and specificity of the model are functions of this cut-off. Sensitivity and specificity can be defined as

$$Sens = \frac{TP}{TP + FN} \tag{30}$$

and

$$Spec = \frac{TN}{TN + FP}, \tag{31}$$

respectively, where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP is the number of false positives. Thus, sensitivity measures the model's ability to correctly detect positive cases ($y_i = 1$, contract terminated) and specificity measures its ability to correctly detect negative cases ($y_i = 0$, contract not terminated). It is desirable to have simultaneously a high sensitivity and a high specificity. The ROC curve is a graph of $Sens$ against $Spec$. Each point on the curve results from a set of model predictions obtained with a unique value of p^*.

Logically, the values of sensitivity and specificity are constrained in $[0, 1]$, and thus, the ROC curve is a curve in $[0, 1]^2$. The higher the area under the curve (AUC), the better the model performance. Thus, AUC values are used to rank the ROC curves between different models. If the AUC is not statistically different and higher than 0.5, it indicates that the model does not perform better than a random classifier [23].

However, it is difficult to turn a modest AUC>0.5 into a successful binary prediction because it is not obvious what cut-off p^* to use in a practical application. In this study, p^* was optimised in the training data, and the binary predictions thus obtained were contrasted against the ground truth. Thus, it was possible to cross tabulate the true label y_i against the predicted label $\hat{y}_i$. This generated 2×2 confusion matrices. In some cases, the observations seemed to concentrate on the diagonal $y_i = \hat{y}_i$. Fisher's exact test [24] was used to test whether this pattern differed significantly from a random assortment.

The performance of NSLM was compared to that of a standard neural network. This neural network was implemented in the widely used Python package keras. Its architecture was chosen so that the model involved two hidden layers with 10 nodes each. A sigmoid activation function was used. The OOS predictive power was measured in the same way as for NSLM.

4 Results

4.1 Univariate analyses

In univariate analysis, most variables were Bonferroni significant and most had a negative effect on the termination probability. The effect sizes were modest, with the largest negative effect observed for jobs in tourism ($E_j = -0.55\%, P < 10^{-7}, P_B < 10^{-5}$) and the largest positive effect for distance to a city ($E_j = 0.35\%, P < 10^{-7}, P_B < 10^{-5}$). With these data, it does not make much point to discuss all effects in detail. Instead, Fig. 4 presents a visual summary of the results. Additionally, Table 1 presents some of the univariate effects of specific interest.

The high statistical significance of the univariate effects may seem enigmatic, but it has a natural explanation: Firstly, the total sample size of the training data was very large ($n = 27,555$). This gave enormous statistical power to the univariate logit regression, even though the effects estimated were small. Secondly, most variables in the geographical, grid-based data were representative of the same socioeconomic phenomena. (Consider e.g. average income of individuals vs. median income of individuals, average income of households and median income of households.) Thus, it was expected that most variables would be either significant or insignificant, and in this case, they were significant.

From Table 1, it may be observed that net migration and net population change had a negative effect on the contract terminations, albeit this effect was not Bonferroni significant, except for the net migration in the training data. Thus, it may be concluded, that the primary hypothesis of the association of the contract terminations and depopulation was partially confirmed. As for the other variables, age of the contract holder, distance to a city and distance to a lake increased the probability of contract termination. All these were highly significant in the training data, but not in the validation data. This probably results from the much smaller number of terminated contracts in the validation data.

Finally, Fig. 5 presents the effect size in validation data against the effect size in training data. From Fig. 5, it may be observed that most variables retained their signs, and the effect sizes correlated moderately ($r = 0.46, P < 10^{-5}$). Fifteen of the strongest effects were Bonferroni significant in both data sets. Thus, it may be concluded that logit regression can yield valuable information regarding the pattern of customer behaviour and the results can be validated, statistical power permitting. In more general terms, this implies that there were probably real effects in the data, not just statistical artefacts.

Table 1. A selection of univariate effects of specific interest. These variables were initially supposed to be good explanatory variables.

Variable	**Training data**			**Validation data**		
	E_j (%)	P	P_B	E_j (%)	P	P_B
net migration	-0.32	<0.0001	0.001 **	-0.16	0.02	1.0
net population change	-0.16	0.03	1.0	-0.15	0.04	1.0
age	0.28	$<10^{-5}$	<0.001 ***	0.06	0.34	1.0
distance to a city	0.35	$<10^{-7}$	$<10^{-5}$ ***	0.03	0.58	1.0
distance to a lake	0.23	<0.001	0.08	0.15	0.02	1.0

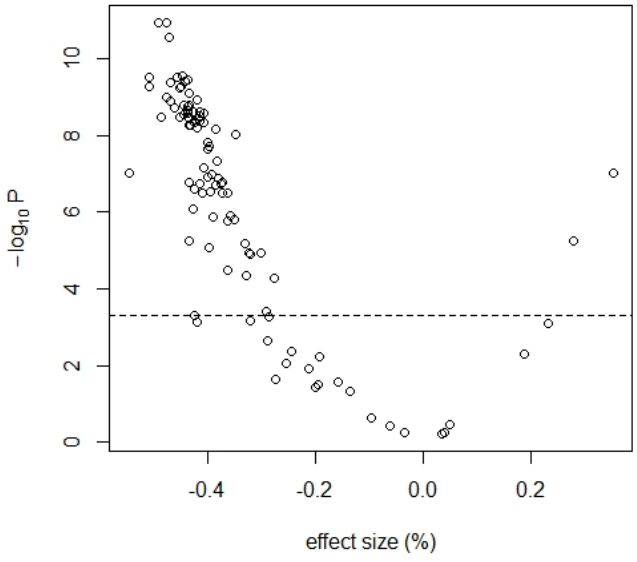

Figure 4. Volcano plot of the univariate analyses. A volcano plot presents the statistical significance against effect size and it is used to illustrate the pattern of significance in massively multivariate data. The dashed line indicates the limit of Bonferroni significance.

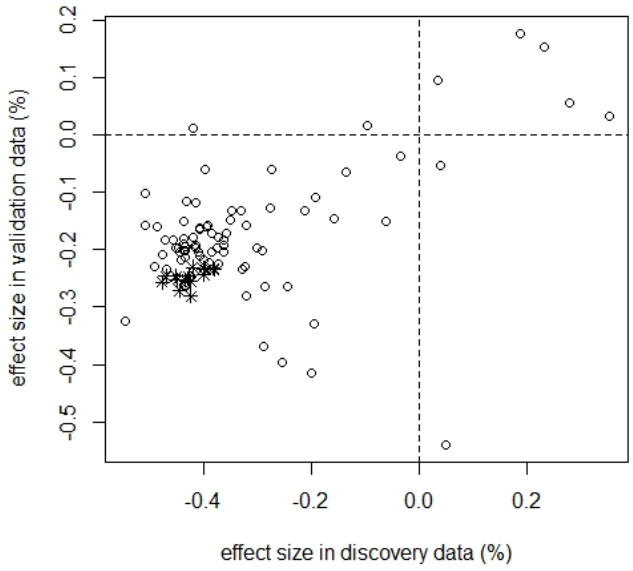

Figure 5. Validation plot of the univariate effects. The effect sizes in validation data are plotted against the effect sizes in training data. Ideally, the points should align around a 45-degree line. Effects which were Bonferroni significant in both data sets are denoted by stars.

4.2 Multivariate analyses

The dimensionality of the data was reduced, as explained in Subsection 3.6. Subsequently, a multivariate model was built by using three methods, i.e. AIC minimisation, crosforw and crosback [13, 19]. Initially, each model choice algorithm was run in the training data. Subsequently, the models thus obtained were re-estimated in the validation data, thus hoping to validate the effects found. Tables 2-4 present the results of these exercises. The P values were not Bonferroni corrected, as this is not customary for multivariate models.

A few general conclusions can be drawn from Tables 2-4. Firstly, all three different model choice algorithms retained age, distance to a city, distance to a lake and income in the model. The first three are individually measured variables, whereas income is a principal component calculated from the geographical data and refers to the income in the grid cell. This suggests the fact that the individually measured covariates are more informative regarding the termination probability, than the PCs calculated from the geographical data. Secondly, the variables retained in the model were usually highly significant in the training data, but less frequently so in the validation data. However, most effects retained their sign in the validation data and in cases where an effect was significant in both data sets, it always retained its sign. Thus, in line with the univariate analyses, it seems that

Table 2. Multivariate model obtained from AIC minimisation. The model was based on logistic regression.

Variable	**Training data**		**Validation data**	
	E_j (%)	P	E_j (%)	P
age	0.18	<0.001 ***	0.05	0.39
distance to a city	0.24	$<10^{-5}$ ***	-0.02	0.74
distance to a lake	0.17	0.002 **	0.13	0.03 *
income	-0.16	0.007 ***	-0.05	0.40
jobs	-0.42	$<10^{-6}$ ***	-0.25	0.003 **

Table 3. Multivariate model obtained from the crosforw algorithm [13]. The model was based on OLS regression and cross validation.

Variable	**Training data**		**Validation data**	
	E_j (%)	P	E_j (%)	P
age	0.23	<0.001 ***	0.15	0.02 *
distance to a city	0.27	<0.0001 ***	0.06	0.39
distance to a lake	0.20	0.002 **	-0.03	0.67
education	-0.06	0.80	-0.14	0.49
income	-0.19	0.01 *	-0.06	0.46
jobs	-0.31	0.14	-0.08	0.67

there were real effects in the data.

Looking at both univariate and multivariate analyses, the effect sizes (E_js) were small, ranging in fractions of per cent. One is justified to ask, whether effects of this size have any economic impact, whether they are statistically significant or not. The answer seems to be in the negative (but see Discussion). The mean termination probability was 1.30% in the training data and 0.54% in the validation data. The 90^{th} percentiles of the model predictions were 1.88% and 1.97% for these two data sets. In fact, the model predictions were <10% for all contracts. Thus, the logit regressions always indicated the rational guess to be ”does not terminate”, and consequently, they could not be used to predict the termination status. Consequently, it was informative to turn towards the non-standard logit machine and the intuitions obtained by using it.

Table 4. Multivariate model obtained from the crosback algorithm [13]. The model was based on OLS regression and cross validation.

Variable	Training data		Validation data	
	E_j (%)	P	E_j (%)	P
age	0.23	<0.001 ***	0.04	0.50
distance to a city	0.25	<0.0001 ***	-0.05	0.45
distance to a lake	0.21	<0.001 ***	0.17	0.005 **
demography	0.84	0.15	0.71	0.26
income	-0.23	0.003 **	-0.13	0.09
households	-1.15	0.04 *	-0.91	0.13

4.3 Non-standard logit machine

As explained in Subsection 3.7.1, NSLM was calibrated by using two methods, cross validation and OLS. Fig. 6 presents the RSS values obtained from the cross validation. Fig. 6 indicated a rough surface, so that the conclusions regarding the 'right' value of $\boldsymbol{\theta}$ were uncertain.

However, OLS gave values $\hat{\theta}_x = 0.64 \ \ (P < 10^{-7})$, $\hat{\theta}_z = 0.48 \ \ (P < 0.001)$ and $\hat{\theta}_w = -0.003 \ \ (P = 0.98)$, thus indicating that the individually measured covariates ($\boldsymbol{X}$) were the most informative ones, followed by the local covariates ($\boldsymbol{Z}$), whereas the areal covariates ($\boldsymbol{W}$) were not to be trusted. Based on these considerations, $\boldsymbol{\theta} = (0.571, 0.429, 0)$ was chosen as the final configuration of this model. This corresponds to the OLS solution, but scaled to the Δ^2 simplex which is the natural range of variation for $\boldsymbol{\theta}$.

Next, both NSLM and the usual neural network were trained on the training data and a prediction for the validation data was attempted. Fig. 7 presents the ROC curves obtained from this exercise. The ROC curve for the non-standard logit machine corresponded the area under curve (AUC) 0.61 with 95% CI: 0.55-0.67, whereas the ROC curve of the neural network had AUC=0.50 (95% CI: 0.43-0.57). This indicated that the logit machine, but not the neural network, had some predictive power towards the validation data. The confidence intervals were calculated by the method of DeLong [23] by using the R package pROC [25].

However, the predictive power could not be transformed into a successful binary prediction: When the natural cut-off $p^* = 0.5$ was used, both models predicted no contracts to be terminated in the validation data. This could not be helped by optimising the cut-off in training data.

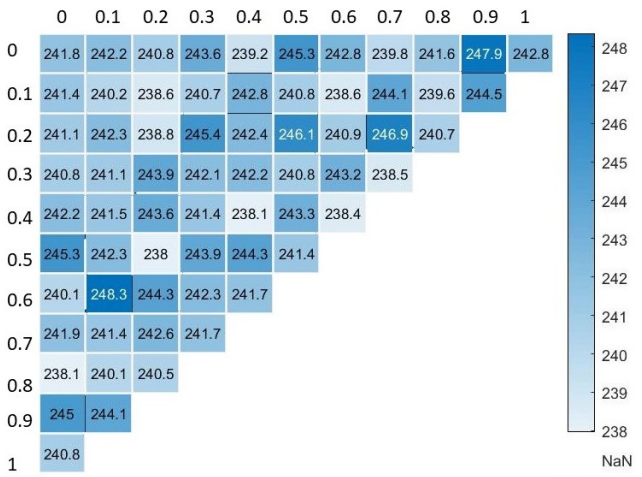

Figure 6. Heatmap of the cross validation for the non-standard logit machine. The figure presents the values of RSS obtained on the Δ^2 simplex for $\boldsymbol{\theta}$. The x axis presents the values of θ_z and the y axis presents the values of θ_x. No consistent pattern is observed.

Finally, it was of interest to see if the uneven distribution of terminated and non-terminated contracts affected the predictive power of these models. To this end, the data were enriched so that the proportion was made even by random sampling (both training data and validation data). The results of this exercise were mixed. In the enriched data, the logit machine had AUC=0.62 (95% CI: 0.51-0.72) and the neural network had AUC=0.50 (95% CI: 0.40-0.59), i.e. the AUC figures were not much improved. The respective ROC curves are presented in Fig. 8.

This time binary prediction was possible, but the predictive power was not statistically significant. With optimised cut-offs, NSLM had an accuracy of 0.56 and the neural network had 0.53. However, when the confusion matrices of Table 5 were tested with Fisher's exact test [24], NSLM had P value 0.26 and neural network had 0.71. This indicated that the models could not be distinguished from a random classifier. With this respect, the machine learning exercise was not successful. However, the ROC curves indicated that NSLM had some predictive power towards the test data, whether enriched or not. This may stem from the fact that NSLM uses context-dependent knowledge, i.e. it makes a distinction between individual, local and areal covariates. To conclude, the performance of NSLM was better than that of the standard neural network.

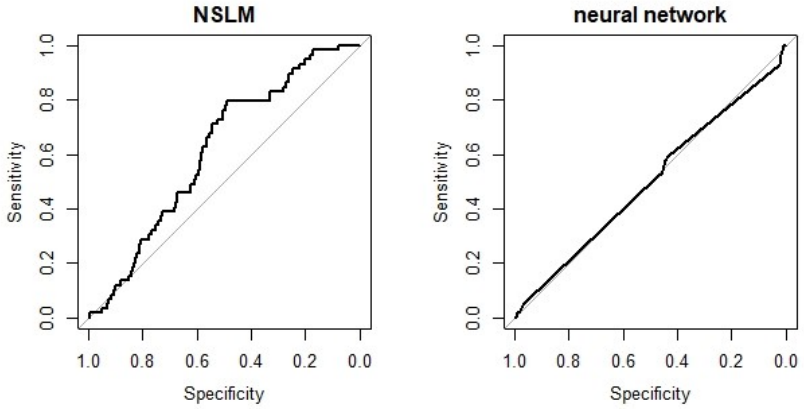

Figure 7. ROC curves for validation data. The models were trained in the training data and predictions were attempted for the validation data. The shape of the ROC curve indicates that the neural network did not perform very well. (Ideally, a model should have simultaneously a high sensitivity and a high specificity.)

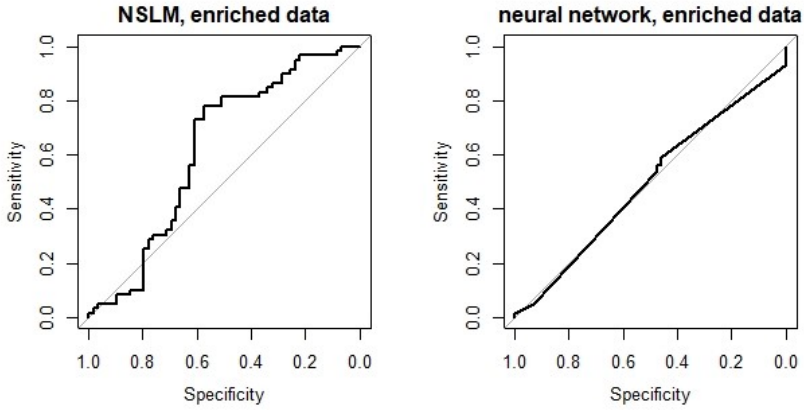

Figure 8. ROC curves for enriched validation data. For this figure, the data were artificially enriched so that the proportion of terminated and non-terminated contracts was even. The models were trained in an enriched training data set and predictions were attempted for enriched validation data.

4.4 Sensitivity analysis

The results of the OOS prediction exercise (Subsection 4.3) were not quite satisfactory. It was hypothesised that perhaps the weak predictive power resulted from some phenomenon idiosyncratic to the test data, North Karelia. As a sensitivity check, the data were re-divided so that North Karelia, South Savo and Kymenlaakso were used as the training data and North Savo was used as the validation data. In this grouping, the training data had 171 terminated contracts and 27,647 non-terminated contracts, whereas the validation data had 263 terminated contracts and 13,457 non-terminated contracts. A few key analyses of the study were re-run on these data.

Table 5. Confusion matrices for enriched test data. The proportion of terminated and non-terminated cases was made artificially even (59 each). NSLM and neural network were trained in the training data (North and South Savo, Kymenlaakso) and the labels of the validation data (North Karelia) were predicted.

True label	**NSLM**		**Neural network**	
	not terminated	terminated	not terminated	terminated
not terminated	37	30	27	24
terminated	22	29	32	35

First, univariate analyses were run for pre-selected variables, as in Subsection 4.1. This resulted in the effect sizes and P values presented in Table 6. This time, no significant associations were found in the training data, but a few were found in the validation data. This probably resulted from the fact that the validation data had more terminated contracts than the training data. Generally, the variables retained their signs.

Second, the principal components were calculated and the AIC-based model choice algorithm was run. Table 7 presents the results of this exercise. In the main analysis (Subsection 4.2), all model choice algorithms had retained the three individual-based variables and the local income level in the model. This time, distance to a city was dropped out from this gold standard, but the other variables and their signs were retained.

Third, the machine learning algorithms were run on these data. The results obtained were similar to those found in the main analysis. In the full, i.e. non-enriched data, NSLM had AUC=0.64 (95% CI: 0.60-0.68) and the neural network had AUC=0.57 (95% CI: 0.54-0.61), thus indicating that there was predictability in the data and both models could detect it. However, when binary prediction was attempted, both models predicted no contracts to be terminated. This could not be helped by optimising the cut-off p^* in the training data.

Thus, an enriched data set was built from these data, as in the main analysis. This data set had 129 terminated and non-terminated contracts in the training data and 175 terminated and non-terminated contracts in the test data, i.e. the frequencies were even. This resulted in the ROC curves presented in Fig. 9. NSLM had AUC=0.62 (95% CI: 0.56-0.68) and the neural network had AUC=0.60 (95% CI: 0.54-0.66). Finally, binary prediction was carried out by using cut-offs optimised in the enriched training data. This resulted in the confusion matrices of Table 8. Based on Table 8, NSLM had accuracy 0.62 and the neural network had 0.55. Fisher's exact test gave $P < 10^{-5}$ for NSLM and $P = 0.04$ for the neural network. This result serves as a proof of concept for the present study: It

Table 6. Univariate effects of specific interest, estimated in an alternative data grouping. This is the direct counterpart of Table 1.

Variable	**Training data**			**Validation data**		
	E_j (%)	P	P_B	E_j (%)	P	P_B
net migration	-0.12	0.13	1.00	-0.40	0.002	0.17
net population change	-0.10	0.05	1.00	-0.25	0.04	1.00
age	0.04	0.35	1.00	0.55	$<10^{-7}$	$<10^{-5}$ ***
distance to a city	0.03	0.54	1.00	0.63	$<10^{-7}$	$<10^{-5}$ ***
distance to a lake	0.08	0.08	1.00	0.64	$<10^{-6}$	<0.0001 ***

Table 7. Multivariate model obtained from AIC minimisation in an alternative data grouping. This is the direct counterpart of Table 2.

Variable	**Training data**		**Validation data**	
	E_j (%)	P	E_j (%)	P
age	0.05	0.31	0.35	<0.0001 ***
distance to a lake	0.10	0.03 *	0.40	$<10^{-5}$ ***
income	-0.08	0.08	-0.31	<0.001 ***
buildings	-0.43	0.01 *	0.25	0.48
jobs	0.08	0.61	-0.70	0.06

is possible to predict the termination status by using the present data and methods, if the data are suitably chosen.

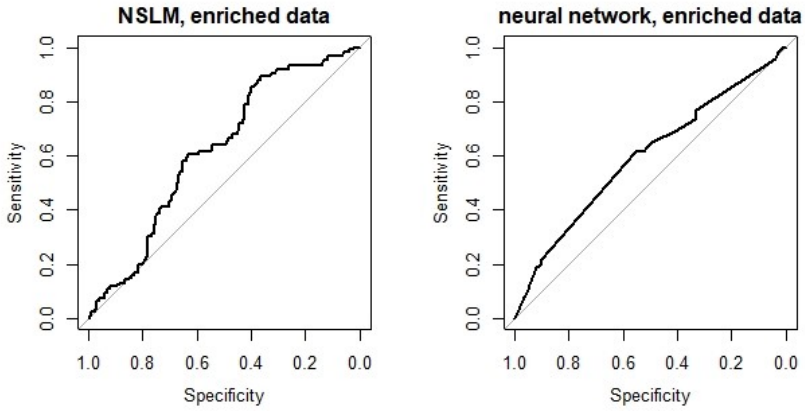

Figure 9. ROC curves for enriched validation data, in an alternative data grouping. For this figure, the data were artificially enriched so that the proportion of terminated and non-terminated contracts was even. This is the counterpart of Fig. 8 in Subsection 4.3.

Table 8. Confusion matrices for enriched test data, in an alternative data grouping. The proportion of terminated and non-terminated cases was made artificially even (175 each). NSLM and neural network were trained in the training data (North Karelia, South Savo and Kymenlaakso) and the labels of the validation data (North Savo) were predicted. The pattern was statistically significant, but it should be kept in mind that the data were specifically hand picked.

True label	**NSLM**		**Neural network**	
	not terminated	terminated	not terminated	terminated
not terminated	72	31	58	40
terminated	103	144	117	135

5 Discussion

5.1 Current study

In this work, data from four DSOs and three government authorities were combined to study the factors affecting the termination of distribution system contracts. Initially, univariate analyses were performed on all variables which yielded highly significant results. Then, a multivariate model was built and machine learning models were trained. Subsection 1.2 characterised the aims of this thesis in terms of three items: 1. discovery of associations, 2. validation of associations and 3. constructing a machine learning model.

Regarding discovery (item 1), associations were certainly found. Most univariate effects were Bonferroni significant, with jobs in tourism and distance to a city having the strongest negative and positive effects, respectively. Net migration, age of the contract holder and distance to a city were among the other significant univariate predictors. The question was not about finding significant factors, but of making use of this knowledge. In multivariate analyses, four variables were always retained in the model: age, distance to a city, distance to a lake and the local income level. Income had a negative effect on the contract terminations, indicating that contracts were terminated less frequently in grid cells with higher income. The other three variables had a positive effect.

Regarding validation (item 2), the results were mixed. It was possible to validate 15 univariate associations, but none of the *a priori* most interesting ones listed in Table 1. Likewise, most multivariate associations could not be reproduced in the validation data, as judged by strict significance criteria. However, the results gave some indication that the associations between contract terminations and the most significant multivariate predictive factors might be real phenomena, and not idiosyncratic to the training data.

Regarding machine learning (item 3), a new type of machine learning algorithm, the NSLM, was constructed. The calibration of the NSLM parameters showed that the individual covariates (age, distance to a city and distance to a lake) were the most important ones, followed by the ones measured from the same 5 km $\times$ 5 km grid cell. Area under the ROC curve showed that NSLM had some predictive power towards the test data.

An alternative geographical grouping of data was used for sensitivity analysis. This approach offered further support for the robust multivariate effects of age, distance to a lake and income. Furthermore, there was OOS predictability in these data: The AUC figures

showed that both methods (NSLM and the neural network) had some predictive power towards the termination status. Moreover, when the data were artificially enriched to have an even number of terminated and non-terminated contracts, binary predictions were statistically significant. This result is best understood as a proof of concept for the present study. However, taken alone, it does not indicate that the present methods have practical utility. This is because enriching the data requires one to know the termination status of each contract – which is precisely the variable one tries to predict.

On the other hand, there is another argument which may demonstrate the usefulness of the present study: As noted in the Introduction, the contract terminations bear economic significance for the DSOs. If there is a substantial chance that the customers will terminate their contracts, a DSO may choose not to provide underground cables to serve these customers. Even a few per cent change in the termination probability might shift the balance in favour of the overhead lines. Thus, it is worthwhile to ask what the models tell about the termination probability.

In these data, with optimal combinations of risk factors, the termination probability of an individual contract ranged from 6% to 8%, whereas the average termination probability was some 1%. However, these were special cases, and the 90^{th} percentiles of the model predictions were 1.88% and 1.97% for the training and validation data sets, respectively. These figures are more representative of the potential economic impact of the present data. It seems unlikely that a termination probability of some 2% in four years can substantially affect the investment decision of a DSO. On the other hand, the probabilities of default encountered in credit risk analysis are often no more substantial, e.g. [26].

5.2 Future study

This work could be extended in a number of ways. Most importantly, the data set included only a few individual-based variables (age of the contract holder, distance to a lake and distance to a city). This was a result of privacy protection regulation: The DSOs had to carefully consider what details of the customers they could hand over to the LUT University. In the present study, only the age of the contract holder and the geographic coordinates of the building were included. (Distance to a lake and distance to a city were calculated from the geographical coordinates.) However, it would be desirable to use more individual-based covariates, as these represent more accurate information regarding the individual customer than the geographical covariates. Age of the building comes to mind.

Finally, in this study, the dimensionality of the primary data (termination + 98 other variables) was reduced by using principal component analysis (resulting in 6 principal components). Subsequently, a model choice was performed between these principal components by using Akaike's information criterion [19] and two cross-validation based algorithms. This is just a scratch into the world of multivariate model building. Other algorithms, such as Tibshirani's LASSO [27] or Zou's Adaptive LASSO [28] could also be tried. In line with this, this work used only two machine learning models, whereas one might consider using other algorithms, such as support vector machines or random forests. It would also be straight forward to extend NSLM to include more levels of variables. This would imply adding more terms to Eq. 25.

It is perhaps an aesthetic flaw that the present study was carried out in three different environments (R, MATLAB and Python), in addition to the geographic information system QGIS. However, it turned out to be convenient to use each programming language/environment for the task it was best suited for. Thus, the bulk of the data handling was carried out in R, whereas MATLAB and Python were used to run computational algorithms (crosforw and crosback [13] and keras, respectively). Building a more unified modelling environment is left for future work.

6 Conclusion

This study identified factors affecting the DSO contract terminations in rural areas of Eastern Finland. It was found that positive net migration and high local income level decreased the risk of contract terminations, whereas age of the contract holder, distance to a city and distance to a lake increased the risk, among other factors. The effects found were small, ranging in fractions of per cent. Strict Bonferroni corrections were used in the univariate analyses and model choice algorithms were employed to construct multivariate models. The results could be partially reproduced in separate validation data. A multivariate regression model gave frequently termination probabilities as high as 1.9% for individual contracts. This may be contrasted with the mean rate of contract terminations found in these data, 1.0%. The machine learning exercise of this work indicated that individually measured covariates were the most informative ones regarding the contract terminations, followed by locally measured ones. The work also demonstrated how difficult it is to predict a rare event, such as a contract termination. Finally, the results of this study emphasise that accurate, individual-specific data are important for customer data analytics.

REFERENCES

[1] J. Haakana. *Impact of Reliability of Supply on Long-Term Development Approaches to Electricity Distribution Networks*. PhD thesis, Lappeenranta University of Technology, Finland, 2013.

[2] K. Knezović, M. Marinelli, A. Zecchino, P. B. Andersen, and C. Traeholt. Supporting involvement of electric vehicles in distribution grids: Lowering the barriers for a proactive integration. *Energy*, 134:458–468, 2017.

[3] Y. Liu, L. Wu, and J. Li. Peer-to-peer (p2p) electricity trading in distribution systems of the future. *The Electricity Journal*, 32(4):2–6, 2019.

[4] V. Marques, N. Bento, and P. M. Costa. The "smart paradox": Stimulate the deployment of smart grids with effective regulatory instruments. *Energy*, 69:96–103, 2014.

[5] Y. Arafat, L. B. Tjernberg, and P. A. Gustafsson. Possibilities of demand side management with smart meters. In *23rd International Conference on Electricity Distribution*. CIRED, 2017.

[6] A. Perosvuo. Maaseudun verkkoliittymäsopimusten asiakaslähtöiseen irtisanomiseen vaikuttavat tekijät pientalo- ja vapaa-ajan asuntokohteissa. Master's thesis, Lappeenranta University of Technology, Finland, 2020.

[7] S. Koskinen, T. Martelin, I-L. Notkola, V. Notkola, K. Pitkänen, M. Jalovaara, E. Mäenpää, A. Ruokolainen, M. Ryynänen, and I. Söderling (Eds.). *Suomen väestö*. Gaudeamus, 2nd edition, 2007.

[8] J. Palkama. The determinants of internal migration in Finland. Master's thesis, Aalto University, Finland, 2018.

[9] N. Kotavaara, O. Kotavaara, J. Rusanen, and T. Muilu. University graduate migration in Finland. *Geoforum*, 96:97–107, 2018.

[10] O. Lehtonen and M. Tykkyläinen. Path dependence in net migration during the ICT boom and two other growth periods: The case of Finland, 1980-2013. *Journal of Evolutionary Economics*, 28(3):547–564, 2018.

[11] M. Vaalavuo and M. W. Sihvola. Are the sick left behind at the peripheries? Health selection in migration to growing urban centres in Finland. *European Journal of Population*, 37(2):341–366, 2021.

[12] C. Adamiak, K. Pitkänen, and O. Lehtonen. Seasonal residence and counterurbanization: The role of second homes in population redistribution in Finland. *GeoJournal*, 82(5):1035–1050, 2017.

[13] H. Haario and V-M. Taavitsainen. *Data analysis toolbox for use with MATLAB™*. ProfMath Oy., Helsinki, Finland, 2004.

[14] I. R. König. Validation in genetic association studies. *Briefings in Bioinformatics*, 12:253–258, 2011.

[15] P. McCullagh and J. A. Nelder. *Generalized linear models*. Routledge, 1st edition, 1983.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[17] A. V. Frane. Are per-family type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods*, 14(1):12–23, 2015.

[18] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[19] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[20] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, 1973.

[21] A. C. Rencher and G. B. Schaalje. *Linear Models in Statistics*. John Wiley & Sons, 2nd edition, 2008.

[22] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[23] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845, 1988.

[24] A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–177, 1992.

[25] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. Müller. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):1–8, 2011.

[26] A. M. Featherstone, L. M. Roessler, and P. J. Barry. Determining the probability of default and risk-rating class for loans in the seventh farm credit district portfolio. *Review of Agricultural Economics*, 28(1):4–23, 2006.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[28] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical association*, 101(476):1418–1429, 2006.

[29] D. Clayton and M. Hills. *Statistical Models in Epidemiology*. Oxford University Press, 1993.

Appendix 1. Inclusion criteria.

Below, there is a list of the inclusion criteria used in this study. A contract had to fulfill all of these criteria to be included in the final sample. It was hoped that by creating a homogeneous data set, successful machine learning and OOS prediction would be possible.

Contracts in rural areas: Urban areas were expected to follow a different dynamic. Thus, data from the Finnish Environment Institute were used to filter out urban areas.

Contracts mapping to small residential buildings: Condominiums were expected to follow a different dynamic. Thus, data from the DSOs were used to filter out condominiums.

Must not be a holiday home: Holiday homes were expected to follow a different dynamic [12]. Thus, data from the DSOs were used to filter out holiday homes.

Must not be possessed by an estate of a deceased person: Inclusion of estates could have biased e.g. the effect of the contract holder's age. Thus, data from the DSOs were used to filter out estates.

Contract termination must have occurred in 1,413 days (if any): Uneven follow-up time is a source of bias [29]. 1,413 days was the feasible maximum in these data. The follow-up time was calculated from the DSO data and unsuitable data points were removed.

Contract termination must have occurred after 2010 (if any): This was justified by the fact that data from 2010 were used as predictive variables and it would be weird to predict past from the future. The contract termination time was observed from the DSO data and unsuitable data points were removed.

Appendix 2. Data sources.

The table below lists the data sources used in this study, together with a categorisation of the covariates. The geographical covariates were measured in 5 km $\times$ 5 km grid cells, perhaps more properly termed as map squares. The variables listed in this table were used directly in the univariate scans (Figs. 4 and 5 in the main matter), whereas the principal components (PCs) were calculated separately for each category of variables and used in the multivariate models (Tables 2-4 in the main matter).

The categories in the data were: A demography, B education, C income, D households, E buildings and F jobs. Specifically, net migration and net population change were left out of these categories because they were based on information observed *after* some of the contract terminations, and thus, they could not be used as predictive factors for the contract terminations. The individually measured variables (termination, age, distance to a city and distance to a lake) were included in the models as they were, i.e. without including them in any PC.

Regarding the data sources listed in Table A1, the following shorthand notation is used for variables which required detailed calculations:

1. calculated from data provided by the DSOs and the National Land Survey of Finland
2. calculated from data provided by the DSOs
3. calculated from data provided by Statistics Finland.

Variables of types 1 and 2 involved geolocation. For variables of type 3, see Subsection 3.2. Additionally, data from the Finnish Environment Institute were used to filter out buildings in the urban areas.

Table A1. List of data sources.

Variable	**Type of measurement**	**Source**	**Category**
contract termination	individual contract	DSOs	-
age (of the contract holder)	individual contract	DSOs	-
distance to a lake	individual building	1.	-
distance to a city	individual building	2.	-
net migration (to the grid cell)	geographical	3.	-

Appendix 2. (continued)

Table A1. List of data sources, cont.

Variable	Level of measurement	Source	Category
net population change	geographical	3.	-
population size	geographical	Statistics Finland	A
women	geographical	Statistics Finland	A
men	geographical	Statistics Finland	A
average age	geographical	Statistics Finland	A
0–2-year-olds	geographical	Statistics Finland	A
3–6-year-olds	geographical	Statistics Finland	A
7–12-year-olds	geographical	Statistics Finland	A
13–15-year-olds	geographical	Statistics Finland	A
16–17-year-olds	geographical	Statistics Finland	A
18–19-year-olds	geographical	Statistics Finland	A
20–24-year-olds	geographical	Statistics Finland	A
25–29-year-olds	geographical	Statistics Finland	A
30–34-year-olds	geographical	Statistics Finland	A
35–39-year-olds	geographical	Statistics Finland	A
40–44-year-olds	geographical	Statistics Finland	A
45–49-year-olds	geographical	Statistics Finland	A
50–54-year-olds	geographical	Statistics Finland	A
55–59-year-olds	geographical	Statistics Finland	A
60–64-year-olds	geographical	Statistics Finland	A
65–69-year-olds	geographical	Statistics Finland	A
70–74-year-olds	geographical	Statistics Finland	A
75–79-year-olds	geographical	Statistics Finland	A
80–84-year-olds	geographical	Statistics Finland	A
85+ year-olds	geographical	Statistics Finland	A
0–17-year-olds	geographical	Statistics Finland	A
people with basic education	geographical	Statistics Finland	B
people with education	geographical	Statistics Finland	B
people with matriculation examination	geographical	Statistics Finland	B
people with vocational training	geographical	Statistics Finland	B
people with bachelor's degree	geographical	Statistics Finland	B

Appendix 2. (continued)

Table A1. List of data sources, cont.

Variable	Level of measurement	Source	Category
people with master's degree	geographical	Statistics Finland	B
average income	geographical	Statistics Finland	C
median income	geographical	Statistics Finland	C
people with low income	geographical	Statistics Finland	C
people with average income	geographical	Statistics Finland	C
people with high income	geographical	Statistics Finland	C
households	geographical	Statistics Finland	D
average household size	geographical	Statistics Finland	D
flat area per inhabitant	geographical	Statistics Finland	D
single households <35-year-old	geographical	Statistics Finland	D
childless households <35-year-old	geographical	Statistics Finland	D
households with children	geographical	Statistics Finland	D
households with children <3-year-old	geographical	Statistics Finland	D
households with children <7-year-old	geographical	Statistics Finland	D
households with children 7–12-year-old	geographical	Statistics Finland	D
households with children 13–17-year-old	geographical	Statistics Finland	D
households with 18–64-year-olds	geographical	Statistics Finland	D
households with 65+ year-olds	geographical	Statistics Finland	D
households owning property	geographical	Statistics Finland	D
tenant households	geographical	Statistics Finland	D
other households	geographical	Statistics Finland	D
average income of households	geographical	Statistics Finland	C
median income of households	geographical	Statistics Finland	C
households with low income	geographical	Statistics Finland	C
households with average income	geographical	Statistics Finland	C
households with high income	geographical	Statistics Finland	C
purchasing power of households	geographical	Statistics Finland	C
holiday homes	geographical	Statistics Finland	E
buildings (excl. holiday homes)	geographical	Statistics Finland	E
other buildings	geographical	Statistics Finland	E
residential buildings	geographical	Statistics Finland	E

Appendix 2. (continued)

Table A1. List of data sources, cont.

Variable	Level of measurement	Source	Category
flats	geographical	Statistics Finland	E
average area of flats	geographical	Statistics Finland	E
flats in small residential buildings	geographical	Statistics Finland	E
flats in apartment buildings	geographical	Statistics Finland	E
jobs	geographical	Statistics Finland	F
jobs in primary production	geographical	Statistics Finland	F
jobs in refinement	geographical	Statistics Finland	F
jobs in services	geographical	Statistics Finland	F
jobs in mining	geographical	Statistics Finland	F
jobs in industry	geographical	Statistics Finland	F
jobs in energy sector	geographical	Statistics Finland	F
jobs in water supply	geographical	Statistics Finland	F
jobs in construction	geographical	Statistics Finland	F
jobs in trade	geographical	Statistics Finland	F
jobs in transport	geographical	Statistics Finland	F
jobs in tourism	geographical	Statistics Finland	F
jobs in ICT	geographical	Statistics Finland	F
jobs in finance	geographical	Statistics Finland	F
jobs in real estate	geographical	Statistics Finland	F
jobs in science and technology	geographical	Statistics Finland	F
jobs in business administration	geographical	Statistics Finland	F
jobs in civil service and defence	geographical	Statistics Finland	F
jobs in education	geographical	Statistics Finland	F
jobs in health care	geographical	Statistics Finland	F
jobs in arts and entertainment	geographical	Statistics Finland	F
jobs in other services	geographical	Statistics Finland	F
jobs in households	geographical	Statistics Finland	F
employed persons	geographical	Statistics Finland	F
unemployed persons	geographical	Statistics Finland	F
students	geographical	Statistics Finland	F
pensioners	geographical	Statistics Finland	F
other occupations	geographical	Statistics Finland	F